Mathematics of Information & Entropy

Volume III: Information Geometry and Statistical Manifolds

The Fisher information metric, exponential families and dual coordinates, dual connections and dually flat spaces, and information projections

---

Every derivation step by step

Every symbol defined

Explicit numerical examples

Geometric and variational intuitions

Silvano Salvador

Mathematics of Information & Entropy Series — Volume III of III

Contents

1.1 Parametric Families as Differentiable Manifolds **7**
From a Family of Distributions to a Manifold . . . 7
Charts, Reparametrisations, and the Role of Coordinates . . . 7
The Tangent Space: Score Functions as Tangent Vectors . . . 8
Smoothness, Regularity, and the Exclusion of Pathological Families . . . 10

1.2 The Fisher Information Metric **10**
Definition: Covariance of the Score . . . 10
Definition via the Hessian of the Log-Likelihood . . . 11
Definition via the Local Quadratic Expansion of KL Divergence . . . 11
Positive Definiteness, Additivity, and Reparametrisation Invariance . . . 12
The Fisher-Rao Distance and Statistical Distinguishability . . . 13

1.3 The Cramér–Rao Bound and Statistical Efficiency **13**
The Scalar Cramér–Rao Bound . . . 14
The Multivariate Cramér–Rao Bound and the Loewner Order . . . 14
Asymptotic Efficiency of the Maximum Likelihood Estimator . . . 15
Geometric Interpretation: Curvature of the Estimator's Distribution . . . 16

1.4 A Minimal Toolkit of Riemannian Geometry **16**
Riemannian Metrics, Lengths, and the Volume Form . . . 16
Geodesics: the Levi-Civita Connection . . . 17
The Riemann Curvature Tensor and Sectional Curvature . . . 18
Pullback Metrics and the Embedding Picture . . . 18
Geodesics in Statistical Manifolds: Explicit Formulae and the Role of Curvature . . . 19

1.5 Canonical Examples: Computing the Fisher Metric **20**
The Bernoulli and Categorical Manifolds . . . 20
The Univariate Gaussian Manifold . . . 21
The Multivariate Gaussian Manifold . . . 22
The Poisson Manifold and a First Look at the General Exponential-Family Formula 23
The Gamma Manifold: the Trigamma Function as a Metric Coefficient . . . 23
The von Mises Manifold: Geometry on a Circular Sample Space . . . 24
A Remark on the Classification of Constant-Curvature Statistical Manifolds . . . 26

1.6 Invariant Measures: the Jeffreys Prior and Reparametrisation **26**
The Parametrisation Problem for "Uninformative" Priors . . . 26
Worked Examples of the Jeffreys Prior . . . 27
Jeffreys Priors, Minimax Codes, and the Bridge to Volume II . . . 29

Numerical Tour: Part 1 — Visualising the Fisher Metric **29**

Formula Reference: Part 1 — Statistical Manifolds and the Fisher Metric **30**

2.1 Exponential Family Structure and Sufficient Statistics **32**
The Canonical Form . . . 32
Sufficiency and the Pitman–Koopman–Darmois Theorem . . . 32
Moments of the Sufficient Statistic via the Log-Partition Function . . . 34

2.2 The Cumulant Generating Function and Legendre Duality **35**

Convex Conjugates: a Brief Recap . . . 35
The Dual Potential is (Minus) the Entropy . . . 35
Worked Dual Potentials . . . 36
The Inverse Hessian Relation . . . 37

2.3 Bregman Divergences and KL Divergence as the Canonical Example 37
Bregman Divergences: Definition and First Properties . . . 37
The Main Theorem: KL Divergence as a Bregman Divergence of the Log-Partition Function . . . 39
The Dual Form: KL as a Bregman Divergence of the Negentropy . . . 39
Why This Matters: Divergence as Geometry, Not Just Algebra . . . 40

2.4 The Generalized Pythagorean Theorem for Bregman Divergences 41
Bregman Projections . . . 41
The Generalized Pythagorean Theorem . . . 42
The Pythagorean Theorem for KL Divergence . . . 44
Looking Ahead: Information Projections and Alternating Minimisation . . . 45

2.5 Canonical Examples: Bregman Divergences and KL Divergence for Exponential Families 46
Categorical: Bregman Divergence of the Negentropy is Literally KL . . . 46
Univariate Gaussian: Recovering the Three-Term KL Formula . . . 47
Poisson: the Generalized KL / Itakura–Saito Connection . . . 48
Multivariate Gaussian: the Trace/Log-Determinant Formula . . . 48
Dirichlet and Multinomial: Dual Coordinates and Conjugacy . . . 49

Numerical Tour: Part 2 — Bregman Divergences, Dual Coordinates, and the Pythagorean Identity 52

Formula Reference: Part 2 — Exponential Families, Dual Coordinates, and Bregman Divergences 53

3.1 General Affine Connections and the Amari–Chentsov Skewness Tensor 54
Beyond Levi-Civita: General Affine Connections . . . 54
The Amari–Chentsov Skewness Tensor . . . 54
The α-Connection Family . . . 55
Dual Connections . . . 56
Curvature of the α-Connections: The Amari–Chentsov Theorem . . . 57

3.2 Exponential Flatness and Mixture Flatness: the e- and m-Connections 59
The e-Connection and its Christoffel Symbols . . . 59
The m-Connection and Mixture Flatness . . . 61
The Meaning of e- and m-Geodesics . . . 62

3.3 The Fundamental Theorem of Dually Flat Spaces 62
Dually Flat Manifolds: Definition and First Properties . . . 63
The Canonical Divergence of a Dually Flat Space . . . 63
The Pythagorean Theorem in Dually Flat Geometry . . . 64
Curvature of the e- and m-Connections vs. Levi-Civita . . . 65
Eguchi's Theorem: Contrast Functions and the Recovery of Dual Geometry . . . 67

3.4 Asymptotic Statistics and the Geometry of Maximum-Likelihood Estimation 68

The MLE as an e-Projection . . . 69
The Cramér–Rao Bound Revisited: Geometric Proof . . . 69
Higher-Order Asymptotics and α-Connection Curvature . . . 70
Natural Gradient Descent . . . 70
Curved Exponential Families and the Geometry of Model Misspecification . . . 71

Numerical Tour: Part 3 — Dual Geodesics, Projections, and Natural Gradient on Bernoulli 74

Formula Reference: Part 3 — Dual Connections, Flat Geometry, and Natural Gradient 75

4.1 Information Projections: e- and m-Projections 76
The Two Fundamental Projections . . . 76
Pythagorean Decomposition and Projection Error . . . 77

4.2 The EM Algorithm as Alternating Projections 77
The EM Algorithm: Setup and Information-Geometric Reformulation . . . 77
Convergence Analysis via the Csiszár–Tusnády Theorem . . . 78
Exponential Family EM: Explicit Formulae . . . 78
Convergence Rate of EM and the Local Geometry at the Fixed Point . . . 79
Generalised EM, ECM, and ECME . . . 80

4.3 Iterative Proportional Fitting and Contingency Tables 81
The IPF Algorithm . . . 81
Convergence Rate of IPF . . . 82
Multi-Marginal IPF and Entropy-Regularised Optimal Transport . . . 83
Generalized Iterative Scaling and Maximum Entropy Duality . . . 84

Numerical Tour: Part 4 — EM on a Gaussian Mixture and IPF on a Contingency Table 86

Formula Reference: Part 4 — Information Projections and EM 87

5.1 The Natural Gradient: Definition and Reparametrisation Invariance 89
Steepest Descent in a Riemannian Manifold . . . 89
Reparametrisation Invariance . . . 89
Amari's Original Derivation and Neural Networks . . . 90

5.2 K-FAC and Practical Natural Gradient 91
Kronecker-Factored Approximate Curvature . . . 91
Damping, Trust Regions, and K-FAC in Practice . . . 92
The Fisher-Rao Gradient Flow . . . 92
Softmax Output Layer: Geometry of Cross-Entropy . . . 93
Mirror Descent and its Relation to Natural Gradient . . . 93

Numerical Tour: Part 5 — Natural Gradient on a Softmax Classifier 95

Formula Reference: Part 5 — Natural Gradient and Practical Optimisation 96

A Notation and Symbols 97

B Riemannian Geometry: Quick Reference 97

C Open Problems and Research Frontiers 99

D Bibliography and Further Reading 100

E Detailed Proofs and Extended Derivations **102**

Closing Synthesis: The Geometry Behind Information Theory **105**

F f-Divergences and Rényi Divergences **106**

G Connections to Volume II: Belief Propagation as Geometric Projection **110**

H Statistical Manifolds in Machine Learning: Extended Examples **112**

I Worked Problems with Full Solutions **117**

Preface to Volume III

Scope and motivation. Volumes I and II treated probability distributions as the objects of study and asked how much information they contain, how efficiently they can be encoded, and how reliably they can be transmitted or compressed. Throughout both volumes, one quantity recurred under many guises: the Kullback–Leibler divergence $D_{\mathrm{KL}}(p \,\|\, q)$, the curvature of the log-partition function, the Fisher information appearing in the Cramér–Rao bound, the Bregman divergence underlying universal coding redundancy. Volume III's premise is that these are not a loose family of analogies but facets of a single mathematical structure: a *family of probability distributions is a Riemannian manifold*, with the Fisher information as its metric, and KL divergence (and its relatives) as a *divergence function* compatible with a pair of mutually dual affine connections on that manifold. This is the subject of information geometry, developed by Amari, Nagaoka, Chentsov, and others from the 1980s onward, and this volume develops it from first principles, with the same standard of explicit, step-by-step derivation and numerical verification as Volumes I and II.

The payoff is twofold. First, *unification*: results that appeared as separate theorems in Volume I (the Cramér–Rao bound, the maximum-entropy/exponential-family duality, the method of types' rate function) and Volume II (the redundancy-capacity theorem, universal coding bounds) are revealed as different shadows of the same geometric structure — the dually flat geometry of exponential families. Second, *computation*: the geometric viewpoint yields algorithms — the natural gradient, the em algorithm, iterative proportional fitting — that are provably correct and often dramatically more efficient than their naive counterparts, precisely because they respect the intrinsic (non-Euclidean) geometry of the underlying statistical manifold.

Structure. The volume is organised into four parts, each building the geometric vocabulary needed for the next.

- **Part 1: Statistical Manifolds and the Fisher Metric.** Parametric families of distributions as differentiable manifolds; the Fisher information metric as a Riemannian metric on parameter space, derived from the covariance of the score; the Cramér–Rao bound as the operational meaning of the Fisher metric; a minimal toolkit of Riemannian geometry (tangent spaces, metrics, geodesics, the volume form); explicit computation of the Fisher metric for the Bernoulli, Gaussian, and Poisson families; and the Jeffreys prior as the reparametrisation-invariant volume form induced by the Fisher metric.
- **Part 2: Exponential Families and Dual Coordinates.** The canonical (natural-parameter) form of exponential families and sufficient statistics; the cumulant generating function $\Phi(\theta)$, its convexity, and Legendre–Fenchel duality with the negative entropy $\Psi(\eta)$; the dual θ/η coordinate systems and the identity $\mathbf{G} = \nabla^2\Phi = (\nabla^2\Psi)^{-1}$; Bregman divergences generated by convex potentials, with KL divergence between exponential-family members as the canonical example; the generalized Pythagorean theorem for Bregman projections, specialising to dual (M- and I-) projections for KL divergence; and worked examples computing the full dual apparatus for the families used throughout the volume.
- **Part 3: Dual Connections, α-Geometry, and Dually Flat Manifolds.** The general theory of affine connections beyond the Riemannian (Levi-Civita) one; dual connections ∇, ∇^* defined by the duality relation with respect to a Riemannian metric; the one-parameter family of α-connections interpolating between the exponential ($\alpha = 1$) and mixture ($\alpha = -1$) connections introduced informally in Volume I; the notion of a dually flat manifold and Amari's theorem characterising exponential families as exactly

the dually flat statistical manifolds; curvature, torsion, and the Codazzi equations governing dual connections; and statistical curvature as a measure of the failure of a model to be an exponential family, with consequences for second-order asymptotic efficiency.

- **Part 4: Information Projections, the em Algorithm, and Applications.** Alternating Bregman projections between a mixture family and an exponential family — the em algorithm of Csiszár and Tusnády — and its convergence proof via the Pythagorean theorems of Part 2; the classical EM algorithm for latent-variable models as a special case; iterative proportional fitting (Sinkhorn iterations) for contingency tables and optimal transport; the natural gradient as steepest descent with respect to the Fisher metric, and its role in modern optimisation; geodesics and the exponential map on statistical manifolds; and a closing synthesis connecting the geometric structures of this volume back to the entropy and divergence quantities of Volume I and the coding-theoretic redundancy bounds of Volume II.

Mathematical prerequisites. The reader is assumed to be familiar with the material of Volume I (entropy, KL divergence, exponential families, the maximum-entropy/Bregman duality of Section 2.6), with multivariable calculus and linear algebra (Hessians, positive-definite matrices, the implicit function theorem), and with the basic vocabulary of differentiable manifolds (tangent spaces, smooth maps, charts) at the level typically covered in a first course on differential geometry; the relevant definitions are nonetheless reviewed as they are needed, so that no separate geometry text is required.

Notation conventions. Throughout this volume:

- $\mathcal{S} = \{p_\theta\}_{\theta \in \Theta}$ denotes a statistical manifold: a parametric family of distributions on a sample space $\mathcal{X}$, regarded as a differentiable manifold with $\theta \mapsto p_\theta$ a coordinate chart.
- $\theta = (\theta^1, \dots, \theta^n)$ denotes natural (exponential) coordinates; $\eta = (\eta_1, \dots, \eta_n)$ denotes expectation (mixture/dual) coordinates, related by $\eta = \nabla\Phi(\theta)$ and $\theta = \nabla\Psi(\eta)$.
- $T(x) = (T_1(x), \dots, T_n(x))$ denotes the vector of sufficient statistics of an exponential family $p_\theta(x) = h(x)\exp(\theta^\top T(x) - \Phi(\theta))$.
- $\mathbf{G}(\theta) = (g_{ij}(\theta))$ denotes the Fisher information matrix, the matrix representation of the Fisher metric in θ-coordinates; $\mathbf{G}(\theta) = \nabla^2\Phi(\theta)$ for exponential families.
- $B_\phi(p : q) = \phi(p) - \phi(q) - \nabla\phi(q)^\top(p - q)$ denotes the Bregman divergence generated by a strictly convex, differentiable ϕ.
- ∇, ∇^* denote a pair of mutually dual affine connections on $\mathcal{S}$ with respect to the Fisher metric; $\nabla^{(\alpha)}$ denotes the α-connection, $\alpha \in [-1, 1]$.
- $T_p\mathcal{M}$ denotes the tangent space to a manifold $\mathcal{M}$ at a point p; $F : \mathcal{M} \to \mathcal{N}$ denotes a smooth map between manifolds, with pullback metric F^*h.
- $D_{\mathrm{KL}}(p \,\|\, q)$, $H(X)$, and all other entropy/divergence notation follow Volume I.

Each section follows a consistent structure: (1) a formal **Definition Box** establishing mathematical objects; (2) rigorous derivations with no steps omitted, often presented as a **Theorem Box** followed by a complete proof; (3) at least one **Example Box** with explicit numerical computations on the running families (Bernoulli, Gaussian, Poisson); and (4) extensions and connections to adjacent theory, both within this volume and back to Volumes I and II.

1.1 Parametric Families as Differentiable Manifolds

From a Family of Distributions to a Manifold

The starting point of information geometry is a deceptively simple shift in perspective. Consider a parametric family of probability distributions

$$\mathcal{S} = \{\, p_\theta(x) \;:\; \theta = (\theta^1, \dots, \theta^n) \in \Theta \subseteq \mathbb{R}^n \,\},$$

where x ranges over a sample space $\mathcal{X}$ (discrete or continuous) and Θ is an open subset of $\mathbb{R}^n$. In classical statistics, Θ is simply a labelling set: each θ names one distribution, and the geometry of Θ as a subset of $\mathbb{R}^n$ — distances, angles, straight lines — has no statistical meaning of its own. Information geometry rejects this view. It treats $\mathcal{S}$ itself as a *differentiable manifold*, with $\theta \mapsto p_\theta$ a coordinate chart, and asks: what is the geometry *induced* on Θ by the fact that its points are probability distributions?

The motivating observation is that not all parametrisations are equal from a statistical point of view. Two points θ and θ' that are close in the Euclidean sense $\|\theta - \theta'\|$ may correspond to distributions p_θ and $p_{\theta'}$ that are either nearly indistinguishable or wildly different, depending on the direction and on where in Θ we are. The Fisher information metric, introduced in Section 1.2, is precisely the Riemannian metric that converts "distance in Θ" into "statistical distinguishability," and it is this metric — not the ambient Euclidean structure of $\mathbb{R}^n$ — that is the correct notion of geometry for $\mathcal{S}$.

Statistical Manifold (Preliminary Definition)

A *statistical manifold* is a parametric family $\mathcal{S} = \{p_\theta : \theta \in \Theta\}$ of probability distributions on a sample space $\mathcal{X}$, together with a diffeomorphism $\theta \mapsto p_\theta$ from an open set $\Theta \subseteq \mathbb{R}^n$ onto $\mathcal{S}$, such that:

1. the map $\theta \mapsto p_\theta(x)$ is smooth (at least C^3) in θ for every fixed x (regularity);
2. $p_\theta(x) > 0$ for all $x \in \mathcal{X}$, $\theta \in \Theta$ (full support, so that $\log p_\theta(x)$ is everywhere defined);
3. differentiation under the integral/sum sign with respect to θ is justified (dominated convergence holds).

The dimension n of Θ is called the dimension of $\mathcal{S}$. Each θ is identified with a point $p \in \mathcal{S}$, and we use p and θ interchangeably for that point.

The word "preliminary" signals that this definition is a working one for the parametric case; a fully coordinate-free definition (an infinite-dimensional manifold of all probability measures equivalent to a reference measure, with $\mathcal{S}$ a finite-dimensional submanifold) exists but is not needed for the constructions of this volume, which are entirely local and coordinate-based in the spirit of Amari's original treatment.

Charts, Reparametrisations, and the Role of Coordinates

A single statistical model can be written using many different parametrisations. The univariate Gaussian family, for instance, can be parametrised by (μ, σ) (mean and standard deviation), by (μ, σ^2) (mean and variance), by (μ, τ) with $\tau = 1/\sigma^2$ (precision), or — as we will see in Part 2 — by the *natural parameters* $(\theta^1, \theta^2) = (\mu/\sigma^2, -1/(2\sigma^2))$. Each of these is a smooth, invertible reparametrisation $\theta \leftrightarrow \theta'$, i.e. a *change of coordinates* on the same underlying manifold $\mathcal{S}$.

The central tenet of information geometry. Geometric objects — tangent vectors, metrics, connections, curvature, geodesic distance — are properties of $\mathcal{S}$ itself, not of any particular coordinate system θ. A quantity that depends on the choice of parametrisation (e.g. the Euclidean distance $\|\theta - \theta'\|$ in an arbitrary parametrisation) is *not* geometric. A quantity that transforms covariantly under reparametrisation (e.g. the Fisher-Rao geodesic distance) *is* geometric. One of the recurring payoffs of this volume is identifying which familiar quantities from statistics and machine learning are geometric in this sense, and which are coordinate artefacts.

Under a smooth reparametrisation $\theta' = \phi(\theta)$ with Jacobian $J^i_{\ j} = \partial\theta'^i/\partial\theta^j$, any tangent vector $v = \sum_i v^i \partial_{\theta^i}$ (Section 1.1.3) transforms as $v'^i = \sum_j J^i_{\ j} v^j$, and any bilinear form such as the Fisher metric g_{ij} (Section 1.2) transforms as

$$g'_{kl} = \sum_{i,j} (J^{-1})^i_{\ k} (J^{-1})^j_{\ l}\, g_{ij}.$$

This is the tensor transformation law; quantities obeying it are called *covariant* (lower indices) or *contravariant* (upper indices) tensors. We will use upper indices for parameters/coordinates and tangent vectors, and lower indices for metric components and differential forms, following the convention of Amari–Nagaoka (2000), which is the standard reference for this material.

The Tangent Space: Score Functions as Tangent Vectors

At each point $p_\theta \in \mathcal{S}$, the *tangent space* $T_{p_\theta}\mathcal{S}$ is, intuitively, the space of "infinitesimal directions of motion" from p_θ to nearby distributions. Formally, a tangent vector is the equivalence class of curves $\theta(t)$ with $\theta(0) = \theta$, where two curves are equivalent if they have the same velocity $\dot\theta(0)$. In coordinates, $T_{p_\theta}\mathcal{S}$ is simply $\mathbb{R}^n$, spanned by the coordinate basis vectors $\partial_1, \dots, \partial_n$ with $\partial_i \triangleq \partial/\partial\theta^i$.

What makes $T_{p_\theta}\mathcal{S}$ a *statistical* object rather than a purely formal one is the following identification. Differentiate $\log p_\theta(x)$ with respect to θ^i:

$$\partial_i \log p_\theta(x) \;=\; \frac{1}{p_\theta(x)}\, \partial_i p_\theta(x).$$

This quantity, as a function of x for fixed θ, is the *score function* associated with parameter θ^i. The collection $(\partial_1 \log p_\theta, \dots, \partial_n \log p_\theta)$ is a vector of random variables (functions of x), and the basis vector $\partial_i \in T_{p_\theta}\mathcal{S}$ is *identified* with the score function $\partial_i \log p_\theta(\cdot)$.

Tangent Vectors as Score Functions

The tangent space $T_{p_\theta}\mathcal{S}$ is identified with the linear span of the score functions $\{\partial_i \log p_\theta(\cdot)\}_{i=1}^n$ inside $L^2(p_\theta)$, the space of functions on $\mathcal{X}$ square-integrable with respect to p_θ. A general tangent vector $A = \sum_i a^i \partial_i$ corresponds to the random variable $A(\cdot) = \sum_i a^i\, \partial_i \log p_\theta(\cdot)$.

This identification is the bridge between differential geometry and statistics: tangent vectors *are* (linear combinations of) score functions, and the inner product on $T_{p_\theta}\mathcal{S}$ that we are about to define (Section 1.2) is simply the $L^2(p_\theta)$ inner product of score functions, i.e. their covariance. A first basic fact, used repeatedly, is that score functions have mean zero.

For each i, $\mathbb{E}_{p_\theta}[\partial_i \log p_\theta(X)] = 0$.

Proof. Using $\partial_i \log p_\theta = (\partial_i p_\theta)/p_\theta$,

$$\mathbb{E}_{p_\theta}[\partial_i \log p_\theta(X)] = \int \frac{\partial_i p_\theta(x)}{p_\theta(x)}\, p_\theta(x)\, dx = \int \partial_i p_\theta(x)\, dx = \partial_i \int p_\theta(x)\, dx = \partial_i(1) = 0,$$

where the interchange of ∂_i and $\int$ is justified by the regularity conditions of Definition 1.1.1, and the final step uses that p_θ integrates to 1 for every θ. ■

This zero-mean property is the reason the score functions span a meaningful inner-product space (rather than, say, requiring a separate centring step): the covariance of the score is simply its second moment, and that second moment is exactly the Fisher information, as we develop next.

Example — The Bernoulli Manifold: a One-Dimensional Statistical Manifold

Let $\mathcal{X} = \{0, 1\}$ and $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ for $\theta \in (0, 1)$. This is a 1-dimensional statistical manifold (a curve, not a surface). The log-likelihood is $\log p_\theta(x) = x \log \theta + (1 - x) \log(1 - \theta)$, and the score is

$$\partial_\theta \log p_\theta(x) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x - \theta}{\theta(1-\theta)}.$$

Its mean is $\mathbb{E}_{p_\theta}\left[\frac{X-\theta}{\theta(1-\theta)}\right] = \frac{\mathbb{E}[X]-\theta}{\theta(1-\theta)} = \frac{\theta-\theta}{\theta(1-\theta)} = 0$, confirming Proposition 1.1.4. The tangent space at each θ is one-dimensional, spanned by this single score function; as θ varies, the score function itself varies (its values at $x = 0$ and $x = 1$ depend on θ), which is why the geometry of this 1-manifold is non-trivial despite its low dimension — we compute its Fisher metric explicitly in Section 1.5.

Example — The Univariate Gaussian Manifold

Let $\mathcal{X} = \mathbb{R}$ and $p_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$. This is a 2-dimensional statistical manifold, with coordinates $\theta = (\mu, \sigma)$. The log-density is

$$\log p_{(\mu,\sigma)}(x) = -\log \sigma - \tfrac{1}{2}\log(2\pi) - \frac{(x-\mu)^2}{2\sigma^2},$$

and the two score functions are

$$\partial_\mu \log p = \frac{x-\mu}{\sigma^2}, \qquad \partial_\sigma \log p = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}.$$

Both have mean zero under $p_{(\mu,\sigma)}$: $\mathbb{E}[\partial_\mu \log p] = \mathbb{E}[X - \mu]/\sigma^2 = 0$, and $\mathbb{E}[\partial_\sigma \log p] = \mathbb{E}[(X-\mu)^2]/\sigma^3 - 1/\sigma = \sigma^2/\sigma^3 - 1/\sigma = 0$ since $\mathbb{E}[(X-\mu)^2] = \sigma^2$. The tangent space at (μ, σ) is the 2-dimensional span of these two score functions inside $L^2(p_{(\mu,\sigma)})$; we will see in Section 1.5 that they are in fact *orthogonal* with respect to the Fisher inner product, with squared norms $1/\sigma^2$ and $2/\sigma^2$ respectively — the famous diagonal Fisher metric of the Gaussian.

Smoothness, Regularity, and the Exclusion of Pathological Families

Not every parametric family is a statistical manifold in the above sense. Three standard pathologies are excluded by the regularity conditions of Definition 1.1.1:

(a) Parameter-dependent support. If the support of p_θ depends on θ — e.g. $p_\theta = \mathrm{Unif}[0, \theta]$ — then $\log p_\theta(x)$ is $-\infty$ outside the support, differentiation under the integral fails at the support boundary, and the Fisher information is either undefined or behaves discontinuously. Such families (called *non-regular*) require a separate theory (their estimators converge at rate $1/n$ rather than $1/\sqrt{n}$, with non-Gaussian limiting distributions) and are excluded from the information-geometric framework of this volume.

(b) Non-identifiability. If $\theta \mapsto p_\theta$ is not injective (two distinct parameter values give the same distribution), the manifold structure degenerates: the "manifold" $\mathcal{S}$ has singular points where the chart fails to be a diffeomorphism, and the Fisher information matrix becomes singular there. Mixture models with label-switching symmetry, and overparametrised neural-network likelihoods, are common sources of non-identifiability; Part 5 returns to the consequences for the natural gradient.

(c) Boundary parameters. If Θ is not open — e.g. a Bernoulli parameter $\theta \in [0, 1]$ including the endpoints — the manifold structure (and the tangent space) degenerates at $\theta \in \{0, 1\}$, where the distribution becomes a point mass and the score function is no longer in L^2. We always work with the *open* parameter set, e.g. $\theta \in (0, 1)$ for Bernoulli, and treat boundary points as limits, not as points of the manifold.

> **Summary of Section 1.1.** A regular parametric family $\mathcal{S} = \{p_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^n$ open, smooth $\theta \mapsto p_\theta(x)$, and everywhere-positive densities is a statistical manifold of dimension n. Its tangent space at θ is identified with the span of the score functions $\partial_i \log p_\theta$ inside $L^2(p_\theta)$, which have zero mean. Reparametrisations $\theta' = \phi(\theta)$ act on tangent vectors and on tensors (such as the Fisher metric, introduced next) via the Jacobian of ϕ, and geometric quantities are precisely those that transform covariantly under such reparametrisations.

1.2 The Fisher Information Metric

Definition: Covariance of the Score

The Fisher information metric assigns to each point $\theta \in \Theta$ an $n \times n$ matrix $\mathbf{G}(\theta) = (g_{ij}(\theta))$, defining an inner product on $T_{p_\theta}\mathcal{S}$. Building on Section 1.1, where tangent vectors are score functions, the natural inner product on $L^2(p_\theta)$ between two zero-mean random variables is their covariance, which for zero-mean variables is simply their expected product.

Fisher Information Metric (Definition 1)

For a regular statistical manifold $\mathcal{S} = \{p_\theta\}$, the *Fisher information matrix* at θ is

$$g_{ij}(\theta) \triangleq \mathbb{E}_{p_\theta}\left[\partial_i \log p_\theta(X)\, \partial_j \log p_\theta(X)\right], \qquad i, j = 1, \dots, n.$$

Equivalently, since the scores have zero mean (Proposition 1.1.4), $g_{ij}(\theta) = \mathrm{Cov}_{p_\theta}(\partial_i \log p_\theta, \partial_j \log p_\theta)$. The matrix $\mathbf{G}(\theta) = (g_{ij}(\theta))$ defines the *Fisher information metric* (or Fisher-Rao metric), a Riemannian metric on $\mathcal{S}$: at each point, an inner product $\langle \partial_i, \partial_j \rangle_{p_\theta} = g_{ij}(\theta)$ on the tangent space.

Definition via the Hessian of the Log-Likelihood

A second, equally standard, characterisation expresses g_{ij} as (minus) the expected Hessian of the log-likelihood. The two definitions coincide under the regularity conditions of Definition 1.1.1, and the equivalence itself is informative.

Under the regularity conditions of Definition 1.1.1,

$$g_{ij}(\theta) = \mathbb{E}_{p_\theta}\left[\partial_i \log p_\theta \, \partial_j \log p_\theta\right] = -\,\mathbb{E}_{p_\theta}\left[\partial_i \partial_j \log p_\theta(X)\right].$$

Proof. Start from $\partial_j \log p_\theta = (\partial_j p_\theta)/p_\theta$ and differentiate again with respect to θ^i, using the quotient rule:

$$\partial_i \partial_j \log p_\theta = \partial_i\left(\frac{\partial_j p_\theta}{p_\theta}\right) = \frac{\partial_i \partial_j p_\theta}{p_\theta} - \frac{(\partial_i p_\theta)(\partial_j p_\theta)}{p_\theta^2} = \frac{\partial_i \partial_j p_\theta}{p_\theta} - (\partial_i \log p_\theta)(\partial_j \log p_\theta).$$

Taking expectations under p_θ and using $\mathbb{E}_{p_\theta}[\cdot] = \int (\cdot)\, p_\theta \, dx$,

$$\mathbb{E}_{p_\theta}\left[\partial_i \partial_j \log p_\theta\right] = \int \partial_i \partial_j p_\theta(x)\, dx - g_{ij}(\theta) = \partial_i \partial_j \int p_\theta(x)\, dx - g_{ij}(\theta) = \partial_i \partial_j (1) - g_{ij}(\theta) = -g_{ij}(\theta),$$

which rearranges to $g_{ij}(\theta) = -\mathbb{E}_{p_\theta}[\partial_i \partial_j \log p_\theta]$, as claimed. ■

The Hessian form $g_{ij} = -\mathbb{E}[\partial_i \partial_j \log p_\theta]$ is the one that appears directly in maximum-likelihood asymptotics: it is (the negative of) the expected curvature of the log-likelihood surface at the true parameter. A sharply peaked log-likelihood (large $|\partial_i \partial_j \log p_\theta|$, i.e. large g_{ij}) means the data are very informative about θ — small changes in θ produce large changes in likelihood, so the MLE is estimated precisely. This is the origin of the name *information*.

Definition via the Local Quadratic Expansion of KL Divergence

The third, and from the information-geometric point of view the most fundamental, definition exhibits g_{ij} as the Hessian of the KL divergence $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'})$ at $\theta' = \theta$. This is the definition that explains the name *Fisher-Rao metric* as a metric "derived from" the KL divergence, and it is the starting point for the divergence geometry of Part 4.

For $\theta' = \theta + d\theta$ with $d\theta \to 0$,

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta + d\theta}) = \tfrac{1}{2} \sum_{i,j} g_{ij}(\theta)\, d\theta^i d\theta^j + O(\|d\theta\|^3).$$

Equivalently, $g_{ij}(\theta) = \partial_{\theta'^i} \partial_{\theta'^j} D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'})\,\big|_{\theta' = \theta}$.

Proof. Write $f(\theta') \triangleq D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = \mathbb{E}_{p_\theta}[\log p_\theta(X) - \log p_{\theta'}(X)]$, a function of θ' with θ fixed. Then $f(\theta) = 0$ (KL divergence of a distribution from itself), and the first-order term vanishes: $\partial_{\theta'^i} f(\theta')\big|_{\theta' = \theta} = -\mathbb{E}_{p_\theta}[\partial_i \log p_\theta(X)] = 0$ by Proposition 1.1.4. For the second-order term,

$$\partial_{\theta'^i} \partial_{\theta'^j} f(\theta')\Big|_{\theta' = \theta} = -\mathbb{E}_{p_\theta}\left[\partial_i \partial_j \log p_\theta(X)\right] = g_{ij}(\theta),$$

using Theorem 1.2.2. A second-order Taylor expansion of f around $\theta' = \theta$ thus reads $f(\theta + d\theta) = 0 + 0 + \frac{1}{2}\sum_{ij} g_{ij}(\theta)\, d\theta^i d\theta^j + O(\|d\theta\|^3)$, which is the claimed expansion. ■

This theorem says: *KL divergence, to second order in a small perturbation, is a quadratic form whose matrix is the Fisher information*. Geometrically, $\sqrt{2D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+d\theta})}$ is, to leading order, the Riemannian distance $\sqrt{\sum_{ij} g_{ij} d\theta^i d\theta^j}$ induced by $\mathbf{G}$. Although $D_{\mathrm{KL}}(\cdot\,\|\,\cdot)$ is not itself a metric (it is asymmetric and does not satisfy the triangle inequality globally), its *infinitesimal* behaviour is symmetric and quadratic, and that infinitesimal behaviour *is* a genuine Riemannian metric. This is the precise sense in which "KL divergence induces the Fisher metric."

Positive Definiteness, Additivity, and Reparametrisation Invariance

Three structural properties make $\mathbf{G}(\theta)$ a legitimate Riemannian metric and a useful statistical object.

$\mathbf{G}(\theta)$ is symmetric positive semi-definite for every θ, since $g_{ij} = \mathrm{Cov}(\partial_i \log p_\theta, \partial_j \log p_\theta)$ is a covariance matrix. Under the identifiability condition of Section 1.1.4(b), $\mathbf{G}(\theta)$ is positive *definite*: a zero eigenvalue would correspond to a direction v along which $\sum_i v^i \partial_i \log p_\theta = 0$ almost surely, i.e. a direction that does not change the distribution to first order, contradicting injectivity of $\theta \mapsto p_\theta$.

If $X_1, \dots, X_n$ are i.i.d. from p_θ, the Fisher information of the joint sample $X_{1:n}$ is n times the single-observation Fisher information: $\mathbf{G}_n(\theta) = n\,\mathbf{G}(\theta)$. More generally, the Fisher information of two independent experiments (possibly with different sample spaces but the same θ) is the sum of their individual Fisher informations.

Proof sketch. The log-likelihood of $X_{1:n}$ is $\sum_{k=1}^n \log p_\theta(X_k)$, a sum of i.i.d. terms with mean-zero score (Proposition 1.1.4) and covariance $\mathbf{G}(\theta)$ each. The score of the joint log-likelihood is the sum of the individual scores, and the covariance of a sum of n i.i.d. zero-mean random vectors is n times the covariance of one term. The two-experiment additivity follows by the same argument applied to the (now non-identically distributed but independent) joint score. ■

Let $\theta' = \phi(\theta)$ be a smooth reparametrisation with Jacobian $J^k_i = \partial \theta'^k / \partial \theta^i$, and let $g'_{kl}(\theta')$ denote the Fisher metric in the θ' coordinates. Then

$$g_{ij}(\theta) = \sum_{k,l} J^k_i \, J^l_j \, g'_{kl}(\phi(\theta)),$$

i.e. $\mathbf{G}$ transforms as a covariant rank-2 tensor (in matrix form, $\mathbf{G}(\theta) = J^\top \mathbf{G}'(\theta') J$).

Proof. By the chain rule, $\partial_{\theta^i} \log p_\theta = \partial_{\theta^i} \log p_{\phi^{-1}(\theta')} = \sum_k J^k_i \, \partial_{\theta'^k} \log p_{\theta'}$ (viewing the density as a function of θ' via the reparametrisation). Hence

$$g_{ij}(\theta) = \mathbb{E}\left[\left(\sum_k J^k_i \partial_{\theta'^k} \log p\right)\left(\sum_l J^l_j \partial_{\theta'^l} \log p\right)\right] = \sum_{k,l} J^k_i J^l_j \, \mathbb{E}[\partial_{\theta'^k} \log p \, \partial_{\theta'^l} \log p] = \sum_{k,l} J^k_i J^l_j \, g'_{kl}. \quad ■$$

This is the key structural fact that justifies calling $\mathbf{G}$ a *metric*: it is not an artefact of the coordinates θ but transforms exactly as a Riemannian metric tensor must. Consequences derived from $\mathbf{G}$ that are themselves tensorial (geodesic distances, scalar curvature, the natural gradient direction as a *vector field*) are therefore genuinely geometric, independent of parametrisation — in sharp contrast to, e.g., the Euclidean gradient $\partial L/\partial\theta$ of a loss function, whose numerical value and direction *do* depend on the parametrisation, a point we return to at length in Part 5.

The Fisher-Rao Distance and Statistical Distinguishability

Once $\mathbf{G}(\theta)$ is known to be a genuine Riemannian metric, the manifold $\mathcal{S}$ inherits a Riemannian distance function: the *Fisher-Rao distance* $d_{FR}(p,q)$ between two distributions $p, q \in \mathcal{S}$ is the infimum, over smooth curves $\gamma : [0,1] \to \mathcal{S}$ with $\gamma(0) = p, \gamma(1) = q$, of the length

$$L[\gamma] = \int_0^1 \sqrt{\sum_{ij} g_{ij}(\gamma(t))\, \dot\gamma^i(t)\dot\gamma^j(t)}\; dt.$$

This distance has a precise statistical meaning: by Theorem 1.2.3, for infinitesimally close p, q, $d_{FR}(p,q) \approx \sqrt{2D_{\mathrm{KL}}(p\,\|\,q)}$, so Fisher-Rao distance measures, to leading order, how distinguishable two distributions are based on a single observation. Cramér–Rao (Section 1.3) makes this precise globally: the Fisher-Rao distance controls how many observations are needed to distinguish p_θ from a nearby $p_{\theta'}$ with a prescribed error probability, via the local quadratic approximation and large-deviations arguments (Chernoff–Stein, Volume I, Part 1).

Example — Fisher-Rao Distance on the Bernoulli Manifold

For the Bernoulli family $p_\theta(x) = \theta^x(1-\theta)^{1-x}$, the Fisher information (computed in Example 1.5.1 below) is $g(\theta) = 1/(\theta(1-\theta))$. The Fisher-Rao distance between θ_0 and θ_1 is

$$d_{FR}(\theta_0,\theta_1) = \int_{\theta_0}^{\theta_1} \sqrt{g(\theta)}\, d\theta = \int_{\theta_0}^{\theta_1} \frac{d\theta}{\sqrt{\theta(1-\theta)}} = 2\left[\arcsin\sqrt{\theta_1} - \arcsin\sqrt{\theta_0}\right].$$

This is the (rescaled) arc length on a circle of radius 1 under the substitution $\theta = \sin^2(\phi/2)$: the Bernoulli manifold, with the Fisher-Rao metric, is isometric to a quarter-circle of radius 2. Two Bernoulli parameters near $\theta = 1/2$ (where $g(\theta)$ is smallest, $g(1/2) = 4$) are statistically *less* distinguishable per unit $|\theta_1 - \theta_0|$ than two parameters near the boundary $\theta \to 0$ or $\theta \to 1$ (where $g(\theta) \to \infty$): a small absolute change in a probability near 0 or 1 is much more informative than the same absolute change near $1/2$, exactly as intuition about likelihood ratios suggests.

Summary of Section 1.2. The Fisher information matrix $g_{ij}(\theta) = \mathbb{E}_{p_\theta}[\partial_i \log p_\theta\, \partial_j \log p_\theta] = -\mathbb{E}_{p_\theta}[\partial_i\partial_j \log p_\theta] = \partial_{\theta'^i}\partial_{\theta'^j} D_{\mathrm{KL}}(p_\theta\,\|\,p_{\theta'})\,|_{\theta'=\theta}$ is positive definite (under identifiability), additive over independent samples/experiments, and transforms as a covariant rank-2 tensor under reparametrisation. It equips the statistical manifold $\mathcal{S}$ with a Riemannian metric — the Fisher-Rao metric — whose induced geodesic distance measures asymptotic statistical distinguishability.

1.3 The Cramér–Rao Bound and Statistical Efficiency

The Scalar Cramér–Rao Bound

The Cramér–Rao bound is the classical inequality that gives the Fisher information its operational meaning: it is the *reciprocal* of the best possible variance of an unbiased estimator. We derive it first in the scalar case via the Cauchy–Schwarz inequality, the same technique used throughout Volume I for entropy inequalities.

Let $T(X)$ be an unbiased estimator of θ, i.e. $\mathbb{E}_{p_\theta}[T(X)] = \theta$ for all θ. Then

$$\operatorname{Var}_{p_\theta}(T(X)) \;\geq\; \frac{1}{g(\theta)},$$

where $g(\theta) = \mathbb{E}_{p_\theta}[(\partial_\theta \log p_\theta(X))^2]$ is the (scalar) Fisher information.

Proof. Differentiate the unbiasedness condition $\mathbb{E}_{p_\theta}[T(X)] = \int T(x) p_\theta(x)\, dx = \theta$ with respect to θ:

$$\partial_\theta \int T(x) p_\theta(x)\, dx = \int T(x)\, \partial_\theta p_\theta(x)\, dx = \int T(x)\, \partial_\theta \log p_\theta(x)\; p_\theta(x)\, dx = \mathbb{E}_{p_\theta}[T(X)\, \partial_\theta \log p_\theta(X)] = 1.$$

Since the score has mean zero (Proposition 1.1.4), $\mathbb{E}[T \cdot \partial_\theta \log p_\theta] = \operatorname{Cov}(T, \partial_\theta \log p_\theta)$, so $\operatorname{Cov}(T, \partial_\theta \log p_\theta) = 1$. By the Cauchy–Schwarz inequality for covariances, $|\operatorname{Cov}(T, \partial_\theta \log p_\theta)|^2 \leq \operatorname{Var}(T) \operatorname{Var}(\partial_\theta \log p_\theta)$, i.e. $1 \leq \operatorname{Var}(T) \cdot g(\theta)$, which rearranges to $\operatorname{Var}(T) \geq 1/g(\theta)$. ■

When is equality attained? Cauchy–Schwarz is an equality iff the two random variables are proportional (almost surely): $T(x) - \theta = c(\theta)\, \partial_\theta \log p_\theta(x)$ for some constant $c(\theta)$ (which must equal $1/g(\theta)$ by taking variances). This is exactly the condition that p_θ belongs to an *exponential family* with sufficient statistic proportional to T (Part 2): the Cramér–Rao bound is attained *exactly*, for all θ simultaneously, precisely for exponential families with T as the natural sufficient statistic. For general families, the bound is attained only asymptotically, by the MLE, as Section 1.3.3 discusses.

The Multivariate Cramér–Rao Bound and the Loewner Order

For vector-valued $\theta \in \mathbb{R}^n$ and an unbiased estimator $\mathbf{T}(X) \in \mathbb{R}^n$ of θ (meaning $\mathbb{E}_{p_\theta}[\mathbf{T}(X)] = \theta$ componentwise), the bound becomes a matrix inequality.

$\operatorname{Cov}_{p_\theta}(\mathbf{T}(X)) \succeq \mathbf{G}(\theta)^{-1}$ in the Loewner (positive semi-definite) order, i.e. $\operatorname{Cov}(\mathbf{T}) - \mathbf{G}(\theta)^{-1}$ is positive semi-definite. Equivalently, for every $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^\top \operatorname{Cov}(\mathbf{T})\mathbf{a} \geq \mathbf{a}^\top \mathbf{G}(\theta)^{-1}\mathbf{a}$.

Proof sketch. Apply the scalar Cramér–Rao bound to the scalar parameter $\psi = \mathbf{a}^\top\theta$ and the scalar unbiased estimator $\mathbf{a}^\top\mathbf{T}(X)$ of ψ. The Fisher information of ψ in terms of $\mathbf{G}(\theta)$ requires the delta-method transformation rule for Fisher information (a special case of Theorem 1.2.4 for linear maps), yielding $g_\psi = (\mathbf{a}^\top \mathbf{G}(\theta)^{-1}\mathbf{a})^{-1}$ when $\mathbf{G}(\theta)$ is invertible. The scalar bound $\operatorname{Var}(\mathbf{a}^\top\mathbf{T}) \geq 1/g_\psi$ becomes $\mathbf{a}^\top \operatorname{Cov}(\mathbf{T})\mathbf{a} \geq \mathbf{a}^\top\mathbf{G}(\theta)^{-1}\mathbf{a}$ for every $\mathbf{a}$, which is precisely the Loewner-order statement. ■

In particular, taking $\mathbf{a} = \mathbf{e}_i$, $\operatorname{Var}(T_i) \geq [\mathbf{G}(\theta)^{-1}]_{ii}$: the variance of each coordinate estimator is bounded by the corresponding diagonal entry of the *inverse* Fisher matrix, not by $1/g_{ii}$ —

off-diagonal Fisher information (correlation between scores of different parameters) generically *increases* the achievable lower bound for each individual coordinate, reflecting that estimating one parameter is made harder by uncertainty about the others (a precise statement of the loss from "nuisance parameters").

Example — Gaussian Mean and Variance: Diagonal Fisher, Decoupled Cramér–Rao Bounds

For $\mathcal{N}(\mu, \sigma^2)$ with natural-looking parametrisation $\theta = (\mu, \sigma^2)$, a direct computation (Section 1.5) gives the Fisher matrix $\mathbf{G}(\theta) = \mathrm{diag}(1/\sigma^2,\, 1/(2\sigma^4))$. Since this is diagonal, its inverse is $\mathrm{diag}(\sigma^2,\, 2\sigma^4)$, and the Cramér–Rao bounds for unbiased estimators of μ and σ^2 decouple: $\mathrm{Var}(\hat{\mu}) \geq \sigma^2$ and $\mathrm{Var}(\widehat{\sigma^2}) \geq 2\sigma^4$ (for a single observation; divide by n for an i.i.d. sample of size n, by additivity, Proposition 1.2.5). The sample mean $\bar{X}$ attains the first bound exactly ($\mathrm{Var}(\bar{X}) = \sigma^2/n$) for every n; the unbiased sample variance $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ has $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$, which exceeds $2\sigma^4/n$ by a factor $n/(n-1) \to 1$: S^2 is asymptotically efficient but not exactly efficient at finite n, illustrating the "asymptotic only" attainment discussed below.

Asymptotic Efficiency of the Maximum Likelihood Estimator

While exact attainment of the Cramér–Rao bound is restricted to exponential families with linear sufficient statistics (Section 1.3.1), a much more general asymptotic statement holds for the maximum likelihood estimator (MLE), under regularity conditions slightly stronger than Definition 1.1.1 (existence of a consistent root of the likelihood equation, dominated third-derivative bounds for the Lindeberg-type CLT argument).

Let $\hat{\theta}_n$ be the MLE based on n i.i.d. observations from p_{θ_0}. Under standard regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0},\, \mathbf{G}(\theta_0)^{-1}) \qquad \text{as } n \to \infty.$$

The asymptotic covariance $\mathbf{G}(\theta_0)^{-1}/n$ equals the Cramér–Rao lower bound for the n-sample experiment (by additivity, $\mathbf{G}_n(\theta_0) = n\mathbf{G}(\theta_0)$, so the bound is $\mathbf{G}_n^{-1} = \mathbf{G}(\theta_0)^{-1}/n$). The MLE is therefore *asymptotically efficient*: it asymptotically attains the Cramér–Rao bound, even when no finite-sample unbiased estimator does.

Proof sketch. A first-order Taylor expansion of the likelihood equation $\sum_{k=1}^{n} \partial_\theta \log p_{\hat{\theta}_n}(X_k) = 0$ around θ_0 gives

$$0 = \sum_k \partial_\theta \log p_{\theta_0}(X_k) + \sum_k \partial_\theta^2 \log p_{\theta_0}(X_k)\,(\hat{\theta}_n - \theta_0) + o_p(\sqrt{n}).$$

Dividing by $-n$ and using the law of large numbers, $\frac{1}{n}\sum_k(-\partial_\theta^2 \log p_{\theta_0}(X_k)) \xrightarrow{p} g(\theta_0)$ (Theorem 1.2.2), so

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx g(\theta_0)^{-1} \cdot \frac{1}{\sqrt{n}} \sum_k \partial_\theta \log p_{\theta_0}(X_k).$$

By the central limit theorem, $\frac{1}{\sqrt{n}}\sum_k \partial_\theta \log p_{\theta_0}(X_k) \xrightarrow{d} \mathcal{N}(0, g(\theta_0))$ (the score terms are i.i.d., mean zero, variance $g(\theta_0)$). Hence $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} g(\theta_0)^{-1}\mathcal{N}(0, g(\theta_0)) = \mathcal{N}(0, g(\theta_0)^{-1})$. The multivariate case follows the identical argument with the matrix $\mathbf{G}(\theta_0)$ in place of $g(\theta_0)$. ■

Geometric Interpretation: Curvature of the Estimator's Distribution

The Cramér–Rao bound has a clean geometric reading in terms of the tangent-space identification of Section 1.1.3. An unbiased estimator $T(X)$ with $\mathbb{E}_{p_\theta}[T] = \theta$ and the score function $\partial_\theta \log p_\theta$ are both elements of $L^2(p_\theta)$; the derivative condition $\partial_\theta \mathbb{E}_{p_\theta}[T] = 1$ derived in the proof of Theorem 1.3.1 says exactly that $T - \theta$ has unit "projection" onto the tangent direction $\partial_\theta \log p_\theta$. Cauchy–Schwarz then states that the squared norm $\mathrm{Var}(T)$ of $T - \theta$ in $L^2(p_\theta)$ is at least the squared norm of its projection divided by $\|\partial_\theta \log p_\theta\|^2 = g(\theta)$, i.e. $\mathrm{Var}(T) \geq 1/g(\theta)$. *The Cramér–Rao bound is the statement that the component of an unbiased estimator's fluctuation lying along the score direction has length at least $1/\sqrt{g(\theta)}$, because that component alone must supply the entire derivative of the expectation.* Estimators that achieve equality have *no* fluctuation orthogonal to the score direction — their entire randomness lies along the single tangent direction $\partial_\theta \log p_\theta$, which is exactly the exponential-family condition revisited in Section 2.1.

This reading also explains why $\mathbf{G}(\theta)^{-1}$, not $\mathbf{G}(\theta)$, is the *natural Riemannian object for estimation error*: if $\mathbf{G}(\theta)$ converts coordinate displacements $d\theta$ into statistical distances, then $\mathbf{G}(\theta)^{-1}$ converts statistical-distance-scaled noise (one unit of "indistinguishability") back into coordinate fluctuations. An estimator's covariance ellipse $\{\delta\theta : \delta\theta^\top \mathbf{G}(\theta)\delta\theta \leq 1\}$ being the Cramér–Rao-optimal ellipse is precisely the statement that the estimator's uncertainty region, measured in the *statistical* (Fisher) metric, is a ball of radius 1 — the smallest possible uncertainty "in information units." This is the first hint that the Fisher metric is not just a convenient bookkeeping device but the metric in which estimation error is naturally isotropic.

Summary of Section 1.3. The Cramér–Rao bound $\mathrm{Var}(T) \geq 1/g(\theta)$ (scalar) or $\mathrm{Cov}(\mathbf{T}) \succeq \mathbf{G}(\theta)^{-1}$ (multivariate) follows from Cauchy–Schwarz applied to the unbiasedness derivative and the score function; equality holds exactly only for linear sufficient statistics of exponential families (Part 2), but the MLE attains the bound *asymptotically* for any regular family, with $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{G}(\theta_0)^{-1})$. Geometrically, $\mathbf{G}(\theta)^{-1}$ is the natural metric for estimation uncertainty, dual to the Fisher metric $\mathbf{G}(\theta)$ on the parameter space itself — a duality that recurs, in sharper algebraic form, throughout Parts 2 and 3.

1.4 A Minimal Toolkit of Riemannian Geometry

Sections 1.1–1.3 introduced statistical manifolds and the Fisher metric using only calculus and probability. To go further — to define geodesics, curvature, and (in Part 3) the dual connections that are the real engine of information geometry — a small amount of Riemannian geometry vocabulary is needed. This section develops exactly what is required, always anchored to the statistical manifold examples already in hand, and defers the full theory of affine connections (beyond the Riemannian/Levi-Civita one) to Part 3.

Riemannian Metrics, Lengths, and the Volume Form

Riemannian Metric

A *Riemannian metric* on a manifold $\mathcal{M}$ is a smooth assignment, to each point $p \in \mathcal{M}$, of a symmetric positive-definite bilinear form $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$. In local coordinates $(\theta^1, \ldots, \theta^n)$, $g_p(\partial_i, \partial_j) = g_{ij}(\theta)$, and the *length* of a tangent vector $v = \sum_i v^i \partial_i$ is $\|v\|_g = \sqrt{\sum_{ij} g_{ij} v^i v^j}$. The length of a curve $\gamma : [a, b] \to \mathcal{M}$ is $L[\gamma] = \int_a^b \|\dot{\gamma}(t)\|_g \, dt$, and the

Riemannian distance $d(p, q)$ is the infimum of $L[\gamma]$ over curves joining p to q.

The Fisher information matrix $\mathbf{G}(\theta) = (g_{ij}(\theta))$ of Section 1.2 is exactly such a g_{ij}; the statistical manifold $\mathcal{S}$ with the Fisher metric is a Riemannian manifold, and d_{FR} (Section 1.2.5) is its Riemannian distance. A Riemannian metric also defines a *volume form*, the natural notion of "area" or "volume" on $\mathcal{M}$ that is consistent with the metric:

$$d\mathrm{Vol}_g(\theta) = \sqrt{\det \mathbf{G}(\theta)}\; d\theta^1 \cdots d\theta^n.$$

For the Fisher metric, $\sqrt{\det \mathbf{G}(\theta)}$ is exactly the density of the *Jeffreys prior* (Section 1.6) — the Riemannian volume form and the Jeffreys prior are the same object, which is why the Jeffreys prior is invariant under reparametrisation (volume forms transform correctly under change of coordinates, by the change-of-variables formula with Jacobian determinant $|\det J|$, exactly cancelling the $|\det J|^2$ from $\det(J^\top \mathbf{G}' J) = (\det J)^2 \det \mathbf{G}'$ up to the square root).

Geodesics: the Levi-Civita Connection

A *geodesic* is a curve that locally minimises length — the Riemannian analogue of a straight line. Geodesics are characterised by a second-order ODE involving the *Christoffel symbols* of the metric, which encode how the basis vectors ∂_i "rotate" as one moves across the manifold.

Christoffel Symbols of the Levi-Civita Connection

For a Riemannian metric g_{ij}, the *Christoffel symbols of the first kind* are

$$\Gamma_{ij,k} \triangleq \tfrac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\right),$$

and of the second kind, $\Gamma^l_{\ ij} \triangleq \sum_k g^{kl}\Gamma_{ij,k}$, where $(g^{kl}) = \mathbf{G}^{-1}$. The Levi-Civita connection is the unique affine connection that is (i) torsion-free ($\Gamma^l_{\ ij} = \Gamma^l_{\ ji}$) and (ii) metric-compatible (parallel transport preserves inner products, equivalently $\nabla g = 0$).

Geodesic Equation

A curve $\theta(t)$ is a geodesic of the Levi-Civita connection iff

$$\ddot{\theta}^k(t) + \sum_{i,j} \Gamma^k_{ij}(\theta(t))\,\dot{\theta}^i(t)\dot{\theta}^j(t) = 0, \qquad k = 1, \dots, n.$$

In flat coordinates (where $\Gamma^k_{\ ij} \equiv 0$, i.e. g_{ij} constant), geodesics are straight lines $\theta(t) = \theta_0 + t\,v$.

Why information geometry needs *more* than the Levi-Civita connection. On a general Riemannian manifold, the Levi-Civita connection is canonical — there is only one torsion-free, metric-compatible connection, by the fundamental theorem of Riemannian geometry. But on a *statistical* manifold, there is an additional structure: the family of α-connections (Part 3), of which the Levi-Civita connection is just the member at $\alpha = 0$. The ± 1-connections (e-connection and m-connection) are generally *not* metric-compatible, are mutually *dual* with respect to $\mathbf{G}$, and — for exponential families — are individually *flat* (zero curvature) even when the Levi-Civita connection is curved. This dual-flatness is the central structural fact of Part 3, and is invisible

if one only ever uses the Levi-Civita connection. The Levi-Civita/$\alpha = 0$ connection nonetheless remains useful as the "unbiased" reference connection, e.g. for defining the Riemannian distance d_{FR} of Section 1.2.5, which uses Levi-Civita geodesics.

The Riemann Curvature Tensor and Sectional Curvature

Curvature measures the failure of parallel transport to be path-independent, or equivalently the second-order deviation of geodesics from straight lines.

Riemann Curvature Tensor

The Riemann curvature tensor of a connection ∇ with Christoffel symbols ${\Gamma^l}_{ij}$ is

$$ {R^l}_{kij} = \partial_i {\Gamma^l}_{jk} - \partial_j {\Gamma^l}_{ik} + \sum_m \left({\Gamma^l}_{im}\Gamma^m_{jk} - {\Gamma^l}_{jm}\Gamma^m_{ik} \right). $$

A connection is *flat* if ${R^l}_{kij} \equiv 0$, equivalently if there exist coordinates in which all Christoffel symbols vanish identically (an "affine" coordinate system).

For a 2-dimensional manifold, all the information in ${R^l}_{kij}$ reduces to a single scalar function, the *Gaussian curvature* $K(\theta)$, via $R_{1212} = K \cdot \det \mathbf{G}$ (with $R_{1212} = \sum_l g_{1l} {R^l}_{212}$). For higher dimensions, the *sectional curvature* $K(\Pi)$ of a 2-plane $\Pi \subset T_p\mathcal{M}$ generalises Gaussian curvature: it is the Gaussian curvature of the surface swept out by geodesics tangent to Π at p.

Example — Constant Curvature of the Bernoulli/Categorical Manifold

The Fisher-Rao metric on the $(K-1)$-dimensional simplex of categorical distributions (K outcomes) is isometric, after the substitution $\theta_k = \phi_k^2$ (mapping probabilities to the square roots of probabilities, which lie on the positive part of a sphere $\sum_k \phi_k^2 = 1$), to (a portion of) the round sphere S^{K-1} of radius 2, which has *constant* positive curvature $K = 1/4$. This is why the Bernoulli Fisher-Rao distance (Example 1.2.6) came out as an arc length on a circle (the $K = 2$ case, S^1 of radius 2): the categorical manifold with the Fisher-Rao metric is, up to the factor of 4 in radius, the positive orthant of a sphere. Constant positive curvature has a direct consequence for natural gradient descent (Part 5): geodesics on the categorical manifold are great-circle arcs, and the "correct" step direction for, e.g., a softmax classifier's output layer bends accordingly — a fact exploited by some natural-gradient implementations that treat softmax outputs specially.

Pullback Metrics and the Embedding Picture

A useful technical device, used repeatedly in worked examples (Section 1.5) and in the information-projection theory of Part 4, is the *pullback metric*: if $F : \mathcal{M} \to \mathcal{N}$ is a smooth map and h is a Riemannian metric on $\mathcal{N}$, the pullback F^*h is the metric on $\mathcal{M}$ defined by $(F^*h)_p(u,v) = h_{F(p)}(dF_p(u), dF_p(v))$. In coordinates, if F has Jacobian J, $(F^*h)_{ij} = \sum_{kl} J^k_i J^l_j h_{kl}$ — exactly the tensor transformation law of Theorem 1.2.4, now read as "the metric on $\mathcal{M}$ induced by mapping into $\mathcal{N}$ and using $\mathcal{N}$'s metric." The square-root embedding $\theta_k \mapsto \phi_k = \sqrt{\theta_k}$ of Example 1.4.3 is precisely a pullback: the Fisher-Rao metric on the simplex is the pullback, under the square-root map, of (a multiple of) the round Euclidean metric on the sphere. This "Fisher metric = pullback of a simple ambient metric

under a nonlinear embedding" picture recurs for every exponential family in a different guise: Part 2 shows that the Fisher metric of an exponential family is the pullback, under the map $\theta \mapsto \eta = \mathbb{E}_\theta[T]$, of the Euclidean metric in the dual coordinates η — but only after the correct (non-affine, in general) identification, which is exactly the content of Legendre duality.

Summary of Section 1.4. A Riemannian metric g_{ij} defines lengths, the volume form $\sqrt{\det g}\, d\theta$ (= Jeffreys prior for the Fisher metric), and — via the Levi-Civita connection's Christoffel symbols Γ^l_{ij} — geodesics and curvature R^l_{kij}. The categorical manifold has constant positive curvature (it is a portion of a round sphere under the square-root embedding). Information geometry's distinctive content, however, lies beyond the Levi-Civita connection: the α-connection family of Part 3, individually flat for exponential families even when Levi-Civita is curved, and related to the Fisher metric by a duality that the pullback/embedding picture begins to suggest.

Geodesics in Statistical Manifolds: Explicit Formulae and the Role of Curvature

The abstract Riemannian machinery of the preceding subsections becomes concrete when one asks: what does a shortest path look like on a statistical manifold? The answer depends sensitively on which connection one uses, and the contrast between the Levi-Civita geodesic and the e- or m-geodesic (to be defined precisely in Part 3) already illuminates the distinction between Riemannian and information geometry.

Levi-Civita geodesics on the Bernoulli manifold. The Bernoulli family $\{p_\theta = \mathrm{Ber}(\theta) : \theta \in (0,1)\}$ has Fisher metric $g(\theta) = 1/[\theta(1-\theta)]$ (a one-dimensional Riemannian manifold). The Levi-Civita geodesic equation reduces (in one dimension) to the ODE $\ddot\theta + \Gamma^1_{11}\dot\theta^2 = 0$, where the single Christoffel symbol is

$$\Gamma^1_{11}(\theta) = \frac{1}{2} g^{11}\partial_1 g_{11} = \frac{\theta(1-\theta)}{2} \cdot \frac{1-2\theta}{\theta^2(1-\theta)^2} = \frac{1-2\theta}{2\theta(1-\theta)}.$$

The geodesic distance between θ_0 and θ_1 is

$$d_{FR}(\theta_0,\theta_1) = \int_{\theta_0}^{\theta_1} \frac{d\theta}{\sqrt{\theta(1-\theta)}} = 2\arcsin(\sqrt{\theta_1}) - 2\arcsin(\sqrt{\theta_0}),$$

the *Fisher-Rao distance* on the Bernoulli manifold, also known as the Hellinger arc-length. The change of variables $\phi = 2\arcsin(\sqrt{\theta})$ (the angular reparametrisation) converts the Bernoulli manifold isometrically to a circle arc of radius 1 with the flat Euclidean metric $d\phi^2$ — the pullback picture of Section 1.4.4 made fully explicit.

Fisher-Rao Geodesics for Gaussian and Categorical Manifolds

Univariate Gaussian. On $\{(\mu,\sigma) : \sigma > 0\}$ with Fisher metric $g_{\mu\mu} = 1/\sigma^2$, $g_{\mu\sigma} = 0$, $g_{\sigma\sigma} = 2/\sigma^2$, the Levi-Civita geodesics are Euclidean lines in the Poincaré upper half-plane $(\mu, \sigma\sqrt{2}) \in \mathbb{H}^2$ with the hyperbolic metric $ds^2 = (d\mu^2 + 2\,d\sigma^2)/\sigma^2$. The sectional curvature is $\kappa = -\frac{1}{2}$ (constant negative curvature).

Categorical/simplex. Via the square-root embedding $\phi_k = \sqrt{\theta_k}$ (sending the interior of the $(K-1)$-simplex to the unit sphere S^{K-1}_+), geodesics correspond to great-circle arcs on S^{K-1}_+ scaled by factor $\frac{1}{2}$: $d_{FR}(p,q) = 2\arccos\big(\sum_k \sqrt{p_k q_k}\big) = 2\arccos(\mathrm{BC}(p,q))$, where $\mathrm{BC}(p,q) = \sum_k \sqrt{p_k q_k}$ is the *Bhattacharyya coefficient*. The geodesic joining p and q on the simplex is the great-circle arc $\gamma(t)_k \propto (\cos(t/2)\sqrt{p_k} + \sin(t/2)\sqrt{q_k})^2$,

$t \in [0, d_{FR}(p, q)]$.

Statistical significance of the geodesic distance. The Fisher-Rao distance $d_{FR}(p, q)$ is the unique (up to scale) metric on $\mathcal{P}(\mathcal{X})$ that is invariant under sufficient statistics, by the Chentsov theorem (1972). Concretely: if $\kappa : \mathcal{X} \to \mathcal{Y}$ is any statistic (measurable function), the induced distributions $\kappa_* p$ and $\kappa_* q$ satisfy $d_{FR}(\kappa_* p, \kappa_* q) \leq d_{FR}(p, q)$, with equality if and only if κ is sufficient for $\{p, q\}$. This monotonicity characterises the Fisher metric as the only Riemannian metric on the statistical manifold that respects the data-processing inequality in the geodesic sense.

Comparison with e-geodesics. While the Levi-Civita geodesic minimises Fisher-Rao arc length, the e-geodesic (exponential-family interpolation) is the *mixture in the natural parameter*: $\theta(t) = (1-t)\theta_0 + t\theta_1$. For the Bernoulli family this gives $p_{\theta(t)} = \text{Ber}((1-t)\theta_0 + t\theta_1)$, a straight line in the natural parameter, whereas the Levi-Civita geodesic curves through the arc $\phi(t) = (1-t)\phi_0 + t\phi_1$ (a straight line in the isometric ϕ coordinate). These two curves coincide only in the limit $\sigma \to 0$ (near-degenerate distributions) or when $\theta_0 = \theta_1$.

The m-geodesic (mixture family interpolation) is the *mixture in the mean parameter*: $\eta(t) = (1-t)\eta_0 + t\eta_1$, i.e. $p_{\eta(t)} = (1-t)p_0 + tp_1$ (a literal probability mixture). For the Bernoulli, $\eta = \theta$, so e- and m-geodesics coincide — a degenerate feature of the one-dimensional case. In higher dimensions (multivariate Gaussian, categorical), all three geodesics (Levi-Civita, e, m) are generically distinct, and the gap between them measures the curvature content of the manifold.

Example — Geodesic Triangle Inequality Failure

The Fisher-Rao distance satisfies the triangle inequality (being a Riemannian geodesic distance), but the KL divergence does not: $D_{\text{KL}}(p \,\|\, r) \not\leq D_{\text{KL}}(p \,\|\, q) + D_{\text{KL}}(q \,\|\, r)$ in general. A concrete counterexample: $p = \text{Ber}(0.01)$, $q = \text{Ber}(0.5)$, $r = \text{Ber}(0.99)$. $D_{\text{KL}}(p \,\|\, r) \approx 4.60$, $D_{\text{KL}}(p \,\|\, q) + D_{\text{KL}}(q \,\|\, r) \approx 1.88 + 1.88 = 3.76 < 4.60$: the "direct" KL exceeds the sum of the "legs." By contrast, $d_{FR}(p, r) = 2\arcsin(\sqrt{0.99}) - 2\arcsin(\sqrt{0.01}) \approx \pi$, $d_{FR}(p, q) + d_{FR}(q, r) \approx \pi/2 + \pi/2 = \pi$, so the Fisher-Rao triangle inequality holds with equality (all three points are collinear on the arc, as $q = \text{Ber}(0.5)$ is the midpoint).

1.5 Canonical Examples: Computing the Fisher Metric

This section computes the Fisher information metric explicitly for the families that will serve as running examples throughout the volume: the Bernoulli/categorical family, the Gaussian family (univariate and multivariate), and the Poisson family, together with the general formula for exponential families that subsumes all of them and is developed fully in Part 2.

The Bernoulli and Categorical Manifolds

Example — Fisher Information of the Bernoulli Family

For $p_\theta(x) = \theta^x(1-\theta)^{1-x}$, $x \in \{0, 1\}$, the score was computed in Example 1.1.5: $\partial_\theta \log p_\theta(x) = (x - \theta)/(\theta(1-\theta))$. The Fisher information is its variance (mean zero):

$$g(\theta) = \text{Var}\left(\frac{X-\theta}{\theta(1-\theta)}\right) = \frac{\text{Var}(X)}{[\theta(1-\theta)]^2} = \frac{\theta(1-\theta)}{[\theta(1-\theta)]^2} = \frac{1}{\theta(1-\theta)}.$$

Cross-checking via the Hessian form: $\log p_\theta(x) = x\log\theta + (1-x)\log(1-\theta)$, $\partial_\theta^2 \log p_\theta(x) = -x/\theta^2 - (1-x)/(1-\theta)^2$, so $-\mathbb{E}[\partial_\theta^2 \log p_\theta] = \theta/\theta^2 + (1-\theta)/(1-\theta)^2 = 1/\theta + 1/(1-\theta) = 1/(\theta(1-\theta))$, matching. The Cramér–Rao bound for an unbiased estimator of θ from one Bernoulli trial is $\mathrm{Var}(\hat{\theta}) \geq \theta(1-\theta)$, attained by $T(X) = X$ itself (since X is an unbiased estimator with exactly this variance) — consistent with Section 1.3.1's remark that Bernoulli is an exponential family with linear sufficient statistic $T(x) = x$, so Cramér–Rao is attained exactly.

Example — Fisher Information Matrix of the Categorical Family

For a categorical distribution over K outcomes with probabilities $(\theta_1, \dots, \theta_{K-1})$ (and $\theta_K = 1 - \sum_{k<K} \theta_k$), write $p_\theta(x = k) = \theta_k$ for $k < K$ and θ_K for $k = K$. Using indicator variables $\mathbb{1}[x = k]$, $\log p_\theta(x) = \sum_{k=1}^{K-1} \mathbb{1}[x = k] \log\theta_k + \mathbb{1}[x = K] \log(1 - \sum_{j<K} \theta_j)$. The score is

$$\partial_{\theta_i} \log p_\theta(x) = \frac{\mathbb{1}[x = i]}{\theta_i} - \frac{\mathbb{1}[x = K]}{\theta_K},$$

and the Fisher matrix entries are

$$g_{ij}(\theta) = \mathrm{Cov}\left(\frac{\mathbb{1}[X = i]}{\theta_i} - \frac{\mathbb{1}[X = K]}{\theta_K},\ \frac{\mathbb{1}[X = j]}{\theta_j} - \frac{\mathbb{1}[X = K]}{\theta_K} \right).$$

Using $\mathrm{Cov}(\mathbb{1}[X = i], \mathbb{1}[X = j]) = \theta_i \delta_{ij} - \theta_i \theta_j$ (multinomial covariance with $n = 1$), a short computation gives the closed form

$$g_{ij}(\theta) = \frac{\delta_{ij}}{\theta_i} + \frac{1}{\theta_K}, \qquad i, j = 1, \dots, K-1,$$

i.e. $\mathbf{G}(\theta) = \mathrm{diag}(1/\theta_1, \dots, 1/\theta_{K-1}) + \frac{1}{\theta_K} \mathbf{1}\mathbf{1}^\top$ (a diagonal matrix plus a rank-one correction coupling all coordinates through the "leftover" category K). For $K = 2$ this reduces to $g_{11} = 1/\theta_1 + 1/\theta_2 = 1/(\theta_1(1-\theta_1))$, recovering Example 1.5.1 with $\theta = \theta_1$.

The Univariate Gaussian Manifold

Example — Fisher Information Matrix of $\mathcal{N}(\mu, \sigma^2)$

Using $\theta = (\mu, \sigma)$ and the scores from Example 1.1.6, $\partial_\mu \log p = (x-\mu)/\sigma^2$ and $\partial_\sigma \log p = (x-\mu)^2/\sigma^3 - 1/\sigma$. We need the second moments of these (both have mean zero). Write $Z = (X-\mu)/\sigma \sim \mathcal{N}(0,1)$.

$$g_{\mu\mu} = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma^2} \right)^2 \right] = \frac{1}{\sigma^4} \mathbb{E}[(X-\mu)^2] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2},$$

$$g_{\sigma\sigma} = \mathbb{E}\left[\left(\frac{(X-\mu)^2}{\sigma^3} - \frac{1}{\sigma} \right)^2 \right] = \frac{1}{\sigma^2} \mathbb{E}\left[\left(Z^2 - 1\right)^2 \right] = \frac{1}{\sigma^2} \mathrm{Var}(Z^2) = \frac{2}{\sigma^2},$$

using $\mathrm{Var}(Z^2) = \mathbb{E}[Z^4] - (\mathbb{E}[Z^2])^2 = 3 - 1 = 2$ for $Z \sim \mathcal{N}(0,1)$. For the cross term,

$$g_{\mu\sigma} = \mathbb{E}\left[\frac{X-\mu}{\sigma^2} \cdot \left(\frac{(X-\mu)^2}{\sigma^3} - \frac{1}{\sigma} \right) \right] = \frac{1}{\sigma^3} \mathbb{E}[Z(Z^2-1)] = \frac{1}{\sigma^3} (\mathbb{E}[Z^3] - \mathbb{E}[Z]) = 0,$$

since all odd moments of $\mathcal{N}(0,1)$ vanish. Hence

$$\mathbf{G}(\mu,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

The metric is *diagonal* but *not* constant: it scales as $1/\sigma^2$, so the manifold $\{(\mu,\sigma) : \sigma > 0\}$ with the Fisher-Rao metric is (a scaled copy of) the hyperbolic half-plane $\mathbb{H}^2$ — a space of constant *negative* curvature $-1/2$, in contrast to the constant positive curvature of the categorical simplex (Example 1.4.3). This is one of the most quoted facts in information geometry: *the univariate Gaussian family, as a Riemannian manifold under the Fisher metric, is hyperbolic space.*

Example — Fisher-Rao Geodesics in the Gaussian Half-Plane

The hyperbolic-plane identification of Example 1.5.3 lets us read off Fisher-Rao geodesic distance directly from the classical hyperbolic distance formula. With the metric $ds^2 = (d\mu^2 + 2\,d\sigma^2)/\sigma^2$, rescaling $\sigma \to \sigma/\sqrt{2}$ gives the standard Poincaré half-plane metric $ds^2 = (du^2 + dv^2)/v^2$ with $u = \mu$, $v = \sigma/\sqrt{2}$, for which the distance between (u_1, v_1) and (u_2, v_2) is $d = \operatorname{arccosh}\!\left(1 + \frac{(u_1-u_2)^2+(v_1-v_2)^2}{2v_1v_2}\right)$. Translating back, the Fisher-Rao distance between $\mathcal{N}(\mu_1,\sigma_1^2)$ and $\mathcal{N}(\mu_2,\sigma_2^2)$ is

$$d_{FR}(\mathcal{N}(\mu_1,\sigma_1^2),\mathcal{N}(\mu_2,\sigma_2^2)) = \sqrt{2}\,\operatorname{arccosh}\!\left(1 + \frac{(\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2}{4\sigma_1\sigma_2}\right)^{1/2}.$$

For two Gaussians with the same variance σ differing only in mean by δ, this gives $d_{FR} \approx |\delta|/\sigma$ for small δ — recovering the intuitive "effect size" (difference in means measured in units of standard deviation) as the leading-order Fisher-Rao distance, and explaining why effect sizes such as Cohen's d are, to leading order, Fisher-Rao distances.

The Multivariate Gaussian Manifold

For $\mathcal{N}(\boldsymbol{\mu}, \square)$ with $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\square \succ 0$ a $d \times d$ covariance matrix, the parameter space has dimension $n = d + \binom{d+1}{2}$ (mean components plus distinct entries of the symmetric covariance matrix). The Fisher information matrix is block-structured.

For $\mathcal{N}(\boldsymbol{\mu}, \square)$, the Fisher information separates into a mean block and a covariance block with no cross terms:

$$g_{\boldsymbol{\mu}_i,\boldsymbol{\mu}_j} = (\square^{-1})_{ij}, \qquad g_{\boldsymbol{\mu}_i,\Sigma_{kl}} = 0, \qquad g_{\Sigma_{ij},\Sigma_{kl}} = \tfrac{1}{2}\left[(\square^{-1})_{ik}(\square^{-1})_{jl} + (\square^{-1})_{il}(\square^{-1})_{jk}\right].$$

In particular, for fixed $\boldsymbol{\mu}$ displacement $d\boldsymbol{\mu}$, the contribution to $B(:)$-type squared Fisher-Rao length is $d\boldsymbol{\mu}^\top \square^{-1} d\boldsymbol{\mu}$ — the squared *Mahalanobis distance*.

Proof sketch. The log-density is $\log p(\mathbf{x}) = -\frac{1}{2}\log\det(2\pi\square) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \square^{-1}(\mathbf{x}-\boldsymbol{\mu})$. The score with respect to $\boldsymbol{\mu}$ is $\square^{-1}(\mathbf{x}-\boldsymbol{\mu})$, a linear function of $\mathbf{x}-\boldsymbol{\mu} \sim \mathcal{N}(0,\square)$; its covariance is $\square^{-1}\square\square^{-1} = \square^{-1}$, giving $g_{\boldsymbol{\mu}\boldsymbol{\mu}} = \square^{-1}$. The score with respect to Σ_{ij} involves $(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top$, which is an *even* function of $\mathbf{x}-\boldsymbol{\mu}$, while the $\boldsymbol{\mu}$-score is *odd*; the covariance of an odd and an even function of a symmetric (Gaussian) random variable vanishes, giving the zero cross-block. The covariance block follows from the fourth-moment (Wick/Isserlis) formula for

Gaussian random variables applied to $(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top$. ■

> **The Mahalanobis distance as Fisher-Rao geometry.** The mean block $g_{\boldsymbol{\mu}\boldsymbol{\mu}} = \square^{-1}$ says that, holding $\square$ fixed, the Fisher-Rao metric on the mean parameters *is* $\square^{-1}$ — the inverse covariance, i.e. the precision matrix. Squared Fisher-Rao distance for a mean perturbation is squared Mahalanobis distance. This single fact underlies an enormous amount of applied statistics: the Mahalanobis distance used in outlier detection, discriminant analysis, and Gaussian process regression is not an ad hoc choice but *is* the Fisher-Rao Riemannian distance on the mean parameters of a Gaussian model with known covariance.

The Poisson Manifold and a First Look at the General Exponential-Family Formula

Example — Fisher Information of the Poisson Family

For $p_\lambda(x) = e^{-\lambda}\lambda^x/x!$, $x \in \{0,1,2,\ldots\}$, $\lambda > 0$, $\log p_\lambda(x) = -\lambda + x\log\lambda - \log x!$, so $\partial_\lambda \log p_\lambda(x) = -1 + x/\lambda = (x-\lambda)/\lambda$. The Fisher information is

$$g(\lambda) = \mathrm{Var}\left(\frac{X-\lambda}{\lambda}\right) = \frac{\mathrm{Var}(X)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

The Fisher-Rao distance is $d_{FR}(\lambda_0,\lambda_1) = \int_{\lambda_0}^{\lambda_1} \lambda^{-1/2} d\lambda = 2(\sqrt{\lambda_1} - \sqrt{\lambda_0})$: Fisher-Rao distance on the Poisson manifold is (twice) the Euclidean distance between square roots of the rates — the geometric basis of the classical variance-stabilising transform $\sqrt{X}$ for Poisson data, which maps the Poisson manifold to a space with (approximately) constant Fisher information 1, i.e. a flat Euclidean metric.

In every example above, the Fisher information turned out to be a *covariance of a sufficient statistic*: $\mathrm{Var}(X)$ for Bernoulli/Poisson, the covariance structure of $((X-\mu),(X-\mu)^2)$-type statistics for the Gaussian. This is not a coincidence. Part 2 shows that for any exponential family $p_\theta(x) \propto \exp(\theta^\top T(x))$, the Fisher information matrix is *exactly* the covariance matrix of the sufficient statistic $T(X)$ under p_θ — a single formula that reproduces every computation in this section as a special case, and that is the gateway to the dual-coordinate structure of Parts 2 and 3. Before turning to that general formula, it is worth working through two further families — the Gamma manifold and the von Mises manifold — whose geometry is less familiar than the Bernoulli/Gaussian/Poisson cases but whose Fisher metrics involve genuinely new special functions and a sample space (the circle) of a fundamentally different topology, both of which will recur as worked examples in later parts.

The Gamma Manifold: the Trigamma Function as a Metric Coefficient

Consider the Gamma family with shape $\alpha > 0$ and rate $\beta > 0$, density

$$p_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0,$$

where $\Gamma(\cdot)$ denotes the ordinary Gamma function (no relation to the Christoffel symbols of Part 3, beyond the unfortunate clash of notation, which we flag once and then suppress by always writing the argument explicitly). The log-likelihood is

$$\log p_{\alpha,\beta}(x) = \alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log x - \beta x,$$

and the score with respect to the two parameters is

$$\frac{\partial \log p}{\partial \alpha} = \log \beta - \ (\alpha) + \log x, \qquad \frac{\partial \log p}{\partial \beta} = \frac{\alpha}{\beta} - x,$$

where $(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. The Fisher information matrix is the covariance of this score vector. Since $\mathrm{Var}(\log X) = \psi_1(\alpha)$ (the *trigamma function*, $\psi_1(\alpha) = '(\alpha)$) is a standard moment identity for the Gamma distribution, and $\mathrm{Var}(X) = \alpha/\beta^2$, $\mathrm{Cov}(\log X, X) = 1/\beta$, one obtains

$$\mathbf{G}(\alpha, \beta) = \begin{pmatrix} \psi_1(\alpha) & -1/\beta \\ -1/\beta & \alpha/\beta^2 \end{pmatrix}.$$

> The appearance of the trigamma function is structurally significant: it is the *only* elementary special function needed for the entire Gamma manifold's curvature, and it appears again, unchanged, as the asymptotic variance of the maximum-likelihood shape estimator (Section 1.3) and in the entropy of the Gamma distribution $H(p_{\alpha,\beta}) = \alpha - \log\beta + \log\Gamma(\alpha) + (1-\alpha)\ (\alpha)$. Such recurrence of a single special function across estimation, geometry, and information theory is typical of exponential families: the cumulant generating function and its derivatives (here built from $\log\Gamma$) generate every quantity of interest, a theme made precise in Part 2.

The determinant is $\det\mathbf{G} = \psi_1(\alpha)\cdot\alpha/\beta^2 - 1/\beta^2 = (\alpha\,\psi_1(\alpha) - 1)/\beta^2$. Since $\psi_1(\alpha) \sim 1/\alpha$ for large α (the digamma function's derivative decays like the reciprocal), the bracket $\alpha\,\psi_1(\alpha) \to 1^+$ as $\alpha \to \infty$, so $\det\mathbf{G} \to 0^+$: the Gamma manifold becomes increasingly degenerate (Fisher-Rao-thin in the shape direction) for large shape parameters, reflecting the fact that a Gamma distribution with large α is close to Gaussian (by the central limit theorem applied to the sum-of-exponentials representation) and the shape parameter becomes statistically harder to pin down relative to the rate. Conversely as $\alpha \to 0^+$, $\psi_1(\alpha) \sim 1/\alpha^2 \to \infty$: the manifold becomes infinitely "tall" in the shape direction near $\alpha = 0$, the boundary at which the Gamma family degenerates into a point mass.

Example — The Exponential Distribution as a Boundary Submanifold

Setting $\alpha = 1$ recovers the exponential distribution $p_\beta(x) = \beta e^{-\beta x}$, a one-dimensional curve inside the two-dimensional Gamma manifold. Restricting $\mathbf{G}(\alpha,\beta)$ to $\alpha = 1$ and projecting onto the β-direction gives the scalar Fisher information $g(\beta) = \alpha/\beta^2|_{\alpha=1} = 1/\beta^2$ for the exponential family — consistent with direct computation from $\log p_\beta(x) = \log\beta - \beta x$, since $\partial_\beta \log p_\beta(x) = 1/\beta - x$ and $\mathrm{Var}(1/\beta - X) = \mathrm{Var}(X) = 1/\beta^2$ (using $\mathrm{Var}(X) = 1/\beta^2$ for $\mathrm{Exp}(\beta)$). This is a first instance of a recurring pattern explored in detail in Part 3 (curved exponential families): fixing one natural parameter of an exponential family as a function of the others carves out a *curved* submanifold, and the submanifold's induced Fisher metric is simply the restriction (pullback) of the ambient Fisher metric — not, in general, the Fisher metric one would compute by treating the submanifold as its own one-parameter family from scratch, although in this particular case (a coordinate slice) the two agree.

The von Mises Manifold: Geometry on a Circular Sample Space

A genuinely different flavour of example arises when the sample space is not $\mathbb{R}$ or $\mathbb{N}$ but the circle $\mathcal{X} = [0, 2\pi)$ (angles, directions, circular time-of-day data, phase variables in signal

processing). The von Mises distribution,

$$p_{\mu,\kappa}(x) = \frac{1}{2\pi I_0(\kappa)}\, e^{\kappa\cos(x-\mu)}, \qquad x \in [0, 2\pi),\ \mu \in [0, 2\pi),\ \kappa \geq 0,$$

is the natural circular analogue of the Gaussian: μ is a mean direction, κ a concentration parameter (playing the role of inverse variance), and I_0 is the modified Bessel function of the first kind of order zero, $I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos t}dt$, which normalises the density. As $\kappa \to \infty$ the von Mises distribution converges (after rescaling $x-\mu$ by $\sqrt{\kappa}$) to a Gaussian on the tangent line at μ, while $\kappa = 0$ gives the uniform distribution on the circle — the two extremes that bookend the family.

The log-likelihood is $\log p_{\mu,\kappa}(x) = \kappa\cos(x-\mu) - \log(2\pi I_0(\kappa))$, with scores

$$\frac{\partial \log p}{\partial \mu} = \kappa\sin(x-\mu), \qquad \frac{\partial \log p}{\partial \kappa} = \cos(x-\mu) - A(\kappa),$$

where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$ and I_1 is the modified Bessel function of order one (using $I_0'(\kappa) = I_1(\kappa)$, a standard Bessel-function identity). Because $\mathbb{E}[\sin(X-\mu)] = 0$ and $\mathbb{E}[\cos(X-\mu)] = A(\kappa)$ by direct integration against the von Mises density (the sine integrates to zero by the symmetry of $\cos(x-\mu)$ about $x=\mu$), the two score components are automatically uncorrelated, $\mathrm{Cov}(\partial_\mu \log p, \partial_\kappa \log p) = \kappa\,\mathbb{E}[\sin(X-\mu)(\cos(X-\mu) - A(\kappa))] = 0$, and the Fisher information matrix is diagonal:

$$\mathbf{G}(\mu,\kappa) = \begin{pmatrix} \kappa\, A(\kappa) & 0 \\ 0 & A'(\kappa) \end{pmatrix},$$

where the (μ,μ) entry follows from $\mathbb{E}[\sin^2(X-\mu)] = 1 - \mathbb{E}[\cos^2(X-\mu)]$ together with a Bessel recurrence identity that simplifies $\kappa^2\mathbb{E}[\sin^2(X-\mu)]$ to $\kappa A(\kappa)$, and the (κ,κ) entry is, by definition, the variance of $\cos(X-\mu)$, equal to the derivative $A'(\kappa) = 1 - A(\kappa)^2 - A(\kappa)/\kappa$ (a known closed form for the derivative of the ratio of Bessel functions).

> Two features distinguish the von Mises manifold from every example so far. First, the metric coefficient $g_{\mu\mu} = \kappa A(\kappa)$ depends on κ but *not* on μ: the manifold is homogeneous under rotation of the mean direction, exactly as one would hope on geometric grounds, since shifting μ is an isometry of the circle. Second, unlike the Gaussian manifold (hyperbolic, constant negative curvature) or the categorical manifold (spherical, constant positive curvature), the von Mises manifold does *not* have constant curvature: its Gaussian curvature is a non-elementary function of κ built from $A(\kappa)$ and its derivatives, reflecting the more intricate analytic structure introduced by the Bessel functions. This is the first example in the volume of a statistical manifold without constant curvature, foreshadowing the general classification discussion below.

As $\kappa \to \infty$, a Taylor expansion of the Bessel-function ratio gives $A(\kappa) \to 1 - \frac{1}{2\kappa} + O(\kappa^{-2})$ and $A'(\kappa) \to \frac{1}{2\kappa} + O(\kappa^{-2})$, so $g_{\mu\mu} \to \kappa$ (information about the mean direction grows linearly in concentration, as expected: a sharply peaked von Mises distribution pins down μ very precisely) while $g_{\kappa\kappa} \to 1/(2\kappa)$ (information about κ itself *shrinks* for large κ, since the difference between, say, $\kappa = 100$ and $\kappa = 101$ has a vanishingly small effect on the shape of an already extremely concentrated distribution). This asymmetric information growth — the mean direction becomes easy to estimate while the concentration becomes hard to estimate, in the high-concentration regime — has practical consequences for circular statistics (directional data in geophysics, bioinformatics's protein backbone dihedral angles, and the

phase-retrieval problems of signal processing) that mirror, in a circular setting, the classical heteroskedastic behaviour of variance estimation discussed for the Gaussian manifold in Section 1.5.2.

A Remark on the Classification of Constant-Curvature Statistical Manifolds

The Bernoulli/categorical manifold (constant positive curvature, a sphere) and the Gaussian manifold (constant negative curvature, hyperbolic space) are not coincidentally simple cases chosen for pedagogical convenience: they are, in a precise sense made rigorous by Eguchi and by Amari, two of the few "universal" constant-curvature statistical manifolds that arise once the relevant symmetry group acts transitively on the parameter manifold. The categorical distribution's $K-1$ simplex of probability vectors carries the Fisher metric, and mapping $\theta_i \mapsto 2\sqrt{\theta_i}$ identifies the simplex isometrically with (a portion of) the positive sector of the sphere S^{K-1} of radius 2 in the Euclidean metric — the square-root reparametrisation already encountered in passing as the variance-stabilising transform for Poisson and Bernoulli data in Sections 1.5.1 and 1.5.4 turns out to be precisely the isometry to constant positive curvature. The Gaussian location-scale manifold, by contrast, is isometric to the hyperbolic plane $\mathbb{H}^2$ with curvature $-1/2$ (the precise constant depends on normalisation conventions for σ), via the classical observation that the Fisher-Rao metric $\mathbf{G} = \mathrm{diag}(1/\sigma^2, 2/\sigma^2)$ is, up to the rescaling $\sigma \mapsto \sigma/\sqrt{2}$, the upper-half-plane metric $ds^2 = (d\mu^2 + d\sigma^2)/\sigma^2$ that defines the Poincaré half-plane model of hyperbolic geometry.

> The von Mises manifold of the previous subsection is a useful counterexample precisely because it does *not* fall into either universality class: it is neither spherical nor hyperbolic, and its curvature genuinely varies with κ. The general lesson, due essentially to Chentsov's uniqueness theorem (Section 1.2.5) together with subsequent curvature computations by several authors, is that constant curvature for a statistical manifold is the exception rather than the rule, special to families (categorical, Gaussian) whose sample space carries an unusually large symmetry group acting transitively on the parameter manifold. Most exponential families — Gamma, von Mises, Beta, Weibull-with-fixed-shape and the many others encountered in applied statistics — have Fisher-Rao curvature that is a nontrivial function of the parameters, to be computed case by case using the machinery of Section 1.4 (the Riemann tensor) or, more efficiently for exponential families specifically, the α-connection curvature formulae of Part 3.

> **Summary of Section 1.5.** Explicit Fisher metrics: Bernoulli $g(\theta) = 1/(\theta(1-\theta))$; categorical $\mathbf{G} = \mathrm{diag}(1/\theta_i) + \theta_K^{-1}\mathbf{1}\mathbf{1}^\top$ (constant positive curvature, a portion of a sphere); univariate Gaussian $\mathbf{G} = \mathrm{diag}(1/\sigma^2, 2/\sigma^2)$ (constant negative curvature, hyperbolic plane); multivariate Gaussian, block-diagonal with mean block $\square^{-1}$ (Mahalanobis distance) and a covariance block from Gaussian fourth moments; Poisson $g(\lambda) = 1/\lambda$ (square-root variance stabilisation). All are special cases of "Fisher information = covariance of sufficient statistic," the exponential-family formula developed in Part 2.

1.6 Invariant Measures: the Jeffreys Prior and Reparametrisation

The Parametrisation Problem for "Uninformative" Priors

In Bayesian inference, a common desideratum for a prior $\pi(\theta)$ intended to represent "prior ignorance" is that it should not privilege any particular parametrisation. The flat (Lebesgue)

prior $\pi(\theta) \propto 1$ fails this immediately: if $\pi(\theta) \propto 1$ is "uninformative" about θ, then under the reparametrisation $\theta' = \theta^3$ (say), the induced prior on θ' is $\pi(\theta') \propto |\partial\theta/\partial\theta'| \propto |\theta'|^{-2/3}$, which is sharply non-uniform — so "uniform in θ" and "uniform in $\theta' = \theta^3$" cannot both represent the same state of ignorance. Some canonical choice, invariant under reparametrisation, is needed if "ignorance" is to be a well-defined notion independent of how we choose to label the parameters. Jeffreys's 1946 proposal — using the Fisher information — is precisely the volume form of Section 1.4.1, and its invariance is exactly the tensorial transformation law of Theorem 1.2.4.

The Jeffreys Prior

The *Jeffreys prior* on a statistical manifold $\mathcal{S} = \{p_\theta : \theta \in \Theta\}$ is

$$\pi_J(\theta) \;\propto\; \sqrt{\det \mathbf{G}(\theta)}.$$

If $\theta' = \phi(\theta)$ is a smooth reparametrisation with Jacobian J, and $\pi_J(\theta) \propto \sqrt{\det \mathbf{G}(\theta)}$, $\pi'_J(\theta') \propto \sqrt{\det \mathbf{G}'(\theta')}$ are the Jeffreys priors in the two coordinate systems, then they agree as measures: $\pi_J(\theta)\, d\theta = \pi'_J(\theta')\, d\theta'$.

Proof. By Theorem 1.2.4, $\mathbf{G}(\theta) = J^\top \mathbf{G}'(\theta') J$, so $\det \mathbf{G}(\theta) = (\det J)^2 \det \mathbf{G}'(\theta')$, hence $\sqrt{\det \mathbf{G}(\theta)} = |\det J| \sqrt{\det \mathbf{G}'(\theta')}$. By the change-of-variables formula, $d\theta' = |\det J|\, d\theta$, so $\pi'_J(\theta')\, d\theta' = \sqrt{\det \mathbf{G}'(\theta')} \cdot |\det J|\, d\theta = \sqrt{\det \mathbf{G}(\theta)}\, d\theta = \pi_J(\theta)\, d\theta$. ■

This is the precise sense in which the Jeffreys prior represents the *same* "ignorance" regardless of parametrisation: integrating any region of $\mathcal{S}$ against π_J gives the same number whether the region is described in θ- or θ'-coordinates. Equivalently, π_J *is* the Riemannian volume form of $(\mathcal{S}, \mathbf{G})$ (Section 1.4.1), and volume forms are coordinate-independent by construction.

Worked Examples of the Jeffreys Prior

Example — Jeffreys Prior for the Bernoulli Family

From Example 1.5.1, $g(\theta) = 1/(\theta(1-\theta))$, so $\pi_J(\theta) \propto [\theta(1-\theta)]^{-1/2}$, the density of $\mathrm{Beta}(1/2, 1/2)$ — the *arcsine distribution*. This is the celebrated Jeffreys prior for a coin-flip probability: it places more mass near $\theta = 0$ and $\theta = 1$ than the uniform prior, reflecting the fact (from Example 1.2.6) that the Fisher-Rao metric is "stretched" near the boundary — a small coordinate interval near 0 or 1 corresponds to a large Fisher-Rao-distance interval, and the Jeffreys prior compensates by assigning it more mass, so that equal *Fisher-Rao* volumes get equal prior mass.

Example — Jeffreys Prior for the Gaussian Mean and the Gaussian Scale

For $\mathcal{N}(\mu, \sigma^2)$ with σ *known*, $\mathbf{G} = 1/\sigma^2$ is constant in μ, so $\pi_J(\mu) \propto 1$: the Jeffreys prior for a Gaussian mean is the familiar flat prior. For $\mathcal{N}(\mu, \sigma^2)$ with μ *known* and σ unknown, $g(\sigma) = 2/\sigma^2$ (Example 1.5.3), so $\pi_J(\sigma) \propto 1/\sigma$ — the scale-invariant prior, flat in $\log \sigma$. Both of these recover the standard "reference priors" of Bayesian statistics, but here they are *derived* rather than postulated: they are simply the square-root-Fisher-information measure in each case, and their familiar forms (flat for location, $1/\sigma$ for

scale) are consequences of the corresponding Fisher metrics being constant (location) or $\propto 1/\sigma^2$ (scale).

Example — Jeffreys Prior for the Poisson Family

For Poi(λ), $\mathbf{G}(\lambda) = 1/\lambda$ (Section 1.5.4), so $\pi_J(\lambda) \propto \lambda^{-1/2}$. This is the (improper) prior flat in $\sqrt{\lambda}$, the variance-stabilising transformation for the Poisson. In Bayesian Poisson regression, this prior corresponds to the conjugate Gamma$(1/2, 0)$ limit (an improper Gamma), and its use can be justified either geometrically (it is the Riemannian volume form of the Fisher metric on $\{p_\lambda : \lambda > 0\}$) or by the minimax coding argument (it minimises the maximum redundancy for universal Poisson codes, in the sense of Section 1.6.3 below). The Jeffreys prior for the Poisson is sometimes written $\pi_J(\lambda) \propto \lambda^{-1/2}$ and sometimes as $\pi_J(n) \propto 1/\sqrt{n}$ for the count n, using the Poisson saddlepoint approximation $p(n) \approx (2\pi n)^{-1/2}$ at $n \approx \lambda$; both forms arise in the MDL literature.

Example — Jeffreys Prior for the Full Two-Parameter Gaussian

For $\mathcal{N}(\mu, \sigma^2)$ with *both* parameters unknown, the Fisher metric is $\mathbf{G}(\mu, \sigma) = \mathrm{diag}(1/\sigma^2, 2/\sigma^2)$ (Section 1.5.2), so

$$\pi_J(\mu, \sigma) \propto \sqrt{\det \mathbf{G}(\mu, \sigma)} = \sqrt{\frac{2}{\sigma^4}} = \frac{\sqrt{2}}{\sigma^2}.$$

This is the standard Jeffreys prior for the Gaussian location-scale family. Note that $\pi_J(\mu, \sigma) \propto 1/\sigma^2$ — not $1/\sigma$ — when both parameters are treated jointly. The Jeffreys *rule* (use $\pi_J \propto \sqrt{\det \mathbf{G}}$ for the joint parameter) differs from the product of the marginal Jeffreys priors ($\pi_J(\mu)\pi_J(\sigma) \propto 1 \cdot 1/\sigma = 1/\sigma$): the joint Jeffreys prior accounts for the off-diagonal coupling between location and scale estimation, even though $\mathbf{G}$ is diagonal here (no coupling at the information level), by using the *product of eigenvalues* $1/\sigma^2 \cdot 2/\sigma^2$, not just the diagonal entries. This is the key difference between the Jeffreys rule and the "independence Jeffreys prior" $\prod_k \sqrt{\mathbf{G}_{kk}}$ (the latter is sometimes preferred in high-dimensional problems to avoid the inconsistency of the Jeffreys rule for grouped parameters, as noted by Bernardo and Berger's reference prior theory).

Example — Jeffreys Prior for the Categorical (K-Simplex) Family

For the categorical distribution on K outcomes with natural parameters $\theta_k = p_k$ (the mean parametrisation, simplex Δ^{K-1}), the Fisher metric restricted to the simplex is $g_{jk} = \delta_{jk}/p_k - 1$ (Proposition 1.5.1, rank-$(K-1)$ positive semidefinite). The restricted determinant (on the $(K-1)$-dimensional manifold) is

$$\det{}' \mathbf{G}(\theta) = \prod_{k=1}^{K} p_k^{-1} \cdot K^{-1} \quad \Rightarrow \quad \pi_J(p_1, \dots, p_K) \propto \prod_{k=1}^{K} p_k^{-1/2},$$

which is the Dirichlet$(1/2, \dots, 1/2)$ distribution on the simplex — the multivariate generalisation of the arcsine (Beta$(1/2, 1/2)$) prior. The total Fisher-Rao volume of the $(K-1)$-simplex (the integral $\int \sqrt{\det{}' \mathbf{G}}\, dp$) is finite and equals $\sqrt{K} \cdot \pi^{(K-1)/2}/\Gamma(K/2)$ (proportional to the surface area of S^{K-1}, via the square-root embedding $\phi_k = 2\sqrt{p_k}$), confirming that the Riemannian volume is finite for the categorical family (unlike the

Gaussian or Poisson, which have infinite volume due to the non-compact parameter space).

Jeffreys Priors, Minimax Codes, and the Bridge to Volume II

The Jeffreys prior also has a purely information-theoretic interpretation, connecting back to the universal coding material of Volume II, Part 3. For a parametric family with n parameters and N observations, the *Bayesian mixture code* with Jeffreys prior, $q(x_{1:N}) = \int p_\theta(x_{1:N})\,\pi_J(\theta)\,d\theta$, achieves, under regularity conditions, the minimax optimal redundancy

$$\sup_\theta \left[-\log q(x_{1:N}) - (-\log p_\theta(x_{1:N}))\right] = \frac{n}{2}\log\frac{N}{2\pi} + \log\int \sqrt{\det \mathbf{G}(\theta)}\,d\theta + o(1),$$

the celebrated Rissanen/Clarke–Barron asymptotic minimax redundancy, whose leading term $\frac{n}{2}\log N$ is the parametric complexity penalty familiar from the minimum description length (MDL) principle (Volume II, Appendix C). The Jeffreys-prior mixture is, in this precise sense, the "best possible" universal code for a parametric family — and the appearance of $\sqrt{\det \mathbf{G}(\theta)}$ in the redundancy formula is exactly the Riemannian volume of the parameter manifold, now playing the role of a normalising constant for a minimax-optimal codelength. This is one of several places in this volume where a purely geometric quantity (here, total Riemannian volume $\int \sqrt{\det \mathbf{G}}\,d\theta$) turns out to control an information-theoretic quantity (minimax redundancy) exactly, with no approximation beyond the $o(1)$ term.

Summary of Section 1.6 and Part 1. The Jeffreys prior $\pi_J(\theta) \propto \sqrt{\det \mathbf{G}(\theta)}$ is the Riemannian volume form of the Fisher metric, and is invariant under reparametrisation by the tensorial transformation law of Theorem 1.2.4. It recovers the arcsine prior for Bernoulli, the flat prior for a Gaussian location, and the $1/\sigma$ prior for a Gaussian scale, and its total integral controls the minimax redundancy of Bayesian mixture codes (a bridge to Volume II's universal coding). Part 1 has established the Fisher metric as the canonical Riemannian structure on a statistical manifold, computed it for the families used throughout the volume, and connected it to estimation theory (Cramér–Rao), Riemannian geometry (geodesics, curvature, constant-curvature examples), and Bayesian/information-theoretic ideas (Jeffreys prior, MDL). Part 2 now turns to the richer *dual-coordinate* structure available specifically for exponential families.

Numerical Tour: Part 1 — Visualising the Fisher Metric

This numerical tour computes and visualises the Fisher information metric for the families of Section 1.5, illustrating the curvature statements of Section 1.4.

1. The Gaussian half-plane. Discretise $(\mu, \sigma) \in [-2, 2] \times [0.2, 2]$ on a grid. At each point, compute $\mathbf{G}(\mu, \sigma) = \mathrm{diag}(1/\sigma^2, 2/\sigma^2)$ and draw the "Fisher disk" $\{(\delta\mu, \delta\sigma) : \delta\mu^2/\sigma^2 + 2\delta\sigma^2/\sigma^2 \le \epsilon^2\}$ — an ellipse whose size shrinks as $\sigma \to 0$ (small σ means small coordinate changes are highly distinguishable) and grows as $\sigma \to \infty$. Tiling the plane with these ellipses at fixed Fisher-radius ϵ reproduces the classic "hyperbolic tiling" picture, visually confirming the constant negative curvature of Example 1.5.3.

2. Geodesics in the Gaussian half-plane. Using the Poincaré half-plane identification of Example 1.5.4 ($u = \mu$, $v = \sigma/\sqrt{2}$), geodesics are semicircles centred on the u-axis (or vertical lines, in the degenerate case of pure scale change). Plot several geodesics connecting

$\mathcal{N}(0,1)$ to $\mathcal{N}(\mu_1, \sigma_1^2)$ for various (μ_1, σ_1), and verify numerically that the arc-length integral $\int \sqrt{2D_{\mathrm{KL}}\big(p_{\gamma(t)} \,\|\, p_{\gamma(t+dt)}\big)}$ along the geodesic matches the closed-form d_{FR} formula of Example 1.5.4 to within numerical integration error.

3. The categorical simplex as a sphere. For $K = 3$ (a 2-simplex), apply the map $\theta_k \mapsto \phi_k = \sqrt{\theta_k}$ and plot the image of the simplex $\{\theta_1 + \theta_2 + \theta_3 = 1, \theta_k \geq 0\}$ under this map. Verify numerically that it lies on the sphere $\phi_1^2 + \phi_2^2 + \phi_3^2 = 1$, and that the pullback of the round metric on this sphere (computed via the Jacobian of $\theta \mapsto \phi$) matches $\mathbf{G}(\theta)$ from Example 1.5.2 up to the factor of 4 noted in Example 1.4.3.

4. Monte Carlo verification of $g_{ij} = \mathrm{Cov}(\partial_i \log p, \partial_j \log p)$. For each family, sample $N = 10^5$ points from p_θ, compute the empirical covariance of the score functions, and compare to the closed-form Fisher matrix. This both verifies the formulas of Section 1.5 and illustrates Theorem 1.2.1 directly as a Monte Carlo estimate — the "Fisher information from samples" computation that underlies practical natural-gradient implementations (Part 5), which estimate $\mathbf{G}(\theta)$ from minibatches exactly this way.

5. Jeffreys prior as a histogram reweighting. Sample θ uniformly on $[0,1]$ (Bernoulli parameter) and separately sample $\theta \sim \mathrm{Beta}(1/2, 1/2)$ (Jeffreys). For each sample, draw $n = 20$ Bernoulli trials and compute the posterior. Compare the coverage of equal-tailed 90% credible intervals under the two priors across many repetitions: the Jeffreys-prior posterior achieves closer to nominal frequentist coverage uniformly in the true $\theta \in (0,1)$, including near the boundaries — a numerical illustration of the matching-priors property of Jeffreys priors in one dimension.

Formula Reference: Part 1 — Statistical Manifolds and the Fisher Metric

Score function. $\partial_i \log p_\theta(x)$; mean zero under p_θ.

Fisher information matrix (three equivalent forms).

$$g_{ij}(\theta) = \mathbb{E}_{p_\theta}[\partial_i \log p_\theta \, \partial_j \log p_\theta] = -\mathbb{E}_{p_\theta}[\partial_i \partial_j \log p_\theta] = \partial_{\theta'^i} \partial_{\theta'^j} D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) \Big|_{\theta'=\theta}.$$

Reparametrisation (tensor) law. $\mathbf{G}(\theta) = J^\top \mathbf{G}'(\theta') J$, $J_i^k = \partial \theta'^k / \partial \theta^i$.

Additivity. $\mathbf{G}_n(\theta) = n\mathbf{G}(\theta)$ for n i.i.d. observations.

Cramér–Rao bound. $\mathrm{Var}(T) \geq 1/g(\theta)$ (scalar); $\mathrm{Cov}(\mathbf{T}) \succeq \mathbf{G}(\theta)^{-1}$ (multivariate); equality for linear sufficient statistics of exponential families.

MLE asymptotics. $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{G}(\theta_0)^{-1})$.

Christoffel symbols (Levi-Civita). $\Gamma_{ij,k} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij})$, $\Gamma^l{}_{ij} = \sum_k g^{kl} \Gamma_{ij,k}$.

Geodesic equation. $\ddot{\theta}^k + \sum_{ij} \Gamma^k{}_{ij} \dot{\theta}^i \dot{\theta}^j = 0$.

Jeffreys prior. $\pi_J(\theta) \propto \sqrt{\det \mathbf{G}(\theta)}$, the Riemannian volume form; reparametrisation-invariant.

Canonical Fisher metrics.

- Bernoulli: $g(\theta) = 1/(\theta(1-\theta))$; Jeffreys = $\mathrm{Beta}(1/2, 1/2)$.
- Categorical: $\mathbf{G} = \mathrm{diag}(1/\theta_i) + \theta_K^{-1} \mathbf{1}\mathbf{1}^\top$; isometric to sphere S^{K-1} of radius 2 (constant curvature $1/4$).
- Univariate Gaussian: $\mathbf{G} = \mathrm{diag}(1/\sigma^2, 2/\sigma^2)$; hyperbolic plane (constant curvature $-1/2$); $d_{FR} = \sqrt{2}\,\mathrm{arccosh}(\cdots)^{1/2}$.

- Multivariate Gaussian: mean block $\square^{-1}$ (Mahalanobis), covariance block from Gaussian fourth moments, no cross terms.
- Poisson: $g(\lambda) = 1/\lambda$; $d_{FR} = 2|\sqrt{\lambda_1} - \sqrt{\lambda_0}|$.

Part 2 Exponential Families and Dual Coordinates

Sections 2.1 – 2.5. The canonical (natural-parameter) form of exponential families and sufficient statistics; the cumulant generating function, its convexity, and Legendre–Fenchel duality; the dual θ/η coordinate systems and the identity $\mathbf{G} = \nabla^2\Phi = (\nabla^2\Psi)^{-1}$; Bregman divergences generated by convex potentials, with KL divergence as the canonical example; the generalized Pythagorean theorem for Bregman divergences and its role as the algebraic core of information projections (Part 4).

2.1 Exponential Family Structure and Sufficient Statistics

The Canonical Form

Exponential Family (Natural Form)

A family of distributions $\{p_\theta : \theta \in \Theta\}$ on $\mathcal{X}$ is an *exponential family* in natural (canonical) form if there exist functions $T : \mathcal{X} \to \mathbb{R}^n$ (the *sufficient statistic*), $h : \mathcal{X} \to \mathbb{R}_{>0}$ (the *carrier* or *base* measure), and $\Phi : \Theta \to \mathbb{R}$ (the *log-partition function* or *cumulant generating function*) such that

$$p_\theta(x) = h(x)\,\exp\!\big(\theta^\top T(x) - \Phi(\theta)\big), \qquad \Phi(\theta) = \log \int h(x)\exp(\theta^\top T(x))\,dx.$$

$\theta \in \mathbb{R}^n$ is the *natural* (or *canonical*, or *exponential*) parameter, and $\Theta = \{\theta : \Phi(\theta) < \infty\}$ is the *natural parameter space*, which is always a convex subset of $\mathbb{R}^n$.

Every distribution encountered in Section 1.5 is an exponential family in disguise:

Example — Canonical Forms of the Running Examples

Bernoulli. $p_\theta(x) = \theta^x(1-\theta)^{1-x} = (1-\theta)\exp(x\log\frac{\theta}{1-\theta})$. Setting the *natural parameter* $\eta_{\text{nat}} = \log\frac{\theta}{1-\theta}$ (the log-odds, or logit), $T(x) = x$, $h(x) = 1$, and $\Phi(\eta_{\text{nat}}) = \log(1 + e^{\eta_{\text{nat}}})$ (so that $1-\theta = e^{-\Phi}$), recovers the canonical form $p(x) \propto \exp(\eta_{\text{nat}}x - \Phi(\eta_{\text{nat}}))$.
Univariate Gaussian. $p_{(\mu,\sigma^2)}(x) \propto \exp\!\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) = \exp\!\big(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2}\big)$. This has $T(x) = (x, x^2)$ and natural parameters $\theta = (\theta_1, \theta_2) = \big(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\big)$, so $\sigma^2 = -1/(2\theta_2)$ and $\mu = -\theta_1/(2\theta_2)$. The constraint $\theta_2 < 0$ (i.e. $\sigma^2 > 0$) defines the natural parameter space, a half-plane — the curved boundary of (μ, σ)-space ($\sigma > 0$) has become a *linear* constraint in natural parameters, the first hint of the simplifications to come.
Poisson. $p_\lambda(x) = e^{-\lambda}\lambda^x/x! = \frac{1}{x!}\exp(x\log\lambda - \lambda)$. Natural parameter $\theta = \log\lambda$, $T(x) = x$, $h(x) = 1/x!$, $\Phi(\theta) = \lambda = e^\theta$.
Categorical. With $T(x) = (\mathbb{1}[x=1], \dots, \mathbb{1}[x=K-1])$ and natural parameters $\theta_k = \log(\theta_k^{\text{prob}}/\theta_K^{\text{prob}})$ (log-odds relative to category K), $\Phi(\theta) = \log(1 + \sum_{k<K} e^{\theta_k})$.

Sufficiency and the Pitman–Koopman–Darmois Theorem

The term "sufficient statistic" is not incidental terminology: $T(X)$ is, in the formal statistical sense, *sufficient* for θ — the conditional distribution of X given $T(X)$ does not depend on θ, so $T(X)$ contains all the information in X about θ (Volume I, Part 2, discusses sufficiency from the information-theoretic data-processing perspective: $I(\theta; X) = I(\theta; T(X))$ for any prior on θ, with equality in the data-processing inequality).

Among regular families (Definition 1.1.1) whose support does not depend on θ, the *only* families admitting a sufficient statistic of *fixed* dimension (not growing with sample size N) are exponential families. Equivalently: exponential families are exactly the families for which the dimension of "what the data tell you about θ" does not grow with the amount of data.

This theorem is the conceptual justification for the entire volume's focus on exponential families: they are not an arbitrary special case chosen for mathematical convenience, but (within the regular, fixed-support setting) the *unique* class of models for which a finite-dimensional statistical manifold can summarise an arbitrarily large dataset. The sufficient statistic of the N-sample joint distribution is simply $\sum_{k=1}^{N} T(x_k)$ (or its average), by the additivity of the exponent $\sum_k \theta^\top T(x_k) = \theta^\top \sum_k T(x_k)$ — the entire dataset compresses, losslessly for inference about θ, into the single vector $\sum_k T(x_k) \in \mathbb{R}^n$.

Minimal sufficiency and completeness. A sufficient statistic T is *minimal sufficient* if it is a function of every other sufficient statistic — i.e., it achieves the maximum data compression consistent with retaining all information about θ. For an exponential family $\{p_\theta\}$ with k-dimensional natural parameter θ, the canonical sufficient statistic $T(x) = (T_1(x), \dots, T_k(x))$ is minimal sufficient, and moreover *complete*: for any measurable f, $\mathbb{E}_{p_\theta}[f(T(X))] = 0$ for all θ implies $f = 0$ a.s. Completeness is the statistical property that there are "no nuisance parameters" hiding in T — every function of T that has mean zero for all θ is trivially zero. In geometric terms, the sufficient statistic map $x \mapsto T(x)$ is the projection onto the "information-relevant" component, and completeness says the exponential-family manifold $\{\eta(\theta) : \theta \in \Theta\}$ is not contained in any proper affine subspace of $\mathbb{R}^k$ (the moment map $\eta : \Theta \to \mathbb{R}^k$ has full-rank image).

Let $\hat{\psi}_0$ be any unbiased estimator of $\psi(\theta)$, and let T be a complete sufficient statistic for θ in an exponential family. Then:

1. **Rao-Blackwell.** $\hat{\psi}_{RB} = \mathbb{E}[\hat{\psi}_0 \,|\, T(X)]$ is an unbiased estimator of $\psi(\theta)$ with $\mathrm{Var}_{p_\theta}(\hat{\psi}_{RB}) \le \mathrm{Var}_{p_\theta}(\hat{\psi}_0)$ for all θ (variance improvement by conditioning on T).
2. **Lehmann-Scheffé.** $\hat{\psi}_{RB}$ is the *unique* uniformly minimum variance unbiased estimator (UMVUE) of $\psi(\theta)$.

Geometrically: conditioning on T is an m-projection (Section 4.1) of the estimator's distribution onto the sub-model indexed by T; Rao-Blackwell is the variance-reduction statement for this m-projection, and Lehmann-Scheffé is uniqueness of the UMVUE as the image of this projection, using completeness to rule out alternative projections.

Example — UMVUE for the Poisson Family

For $X_1, \dots, X_N \sim \mathrm{Poi}(\lambda)$ i.i.d., the sufficient statistic is $S_N = \sum_{i=1}^{N} X_i \sim \mathrm{Poi}(N\lambda)$. The UMVUE of $\psi(\lambda) = \lambda^k$ (the k-th power of the rate) is

$$\hat{\psi}_{\mathrm{UMVUE}} = \frac{S_N(S_N-1)\cdots(S_N-k+1)}{N^k} = \frac{(S_N)_k}{N^k},$$

the k-th factorial moment of S_N/N (a descending factorial divided by N^k), derivable

by computing $\mathbb{E}[S_N^k/N^k] \neq \lambda^k$ (biased), then Rao-Blackwellising the indicator-based estimator $\widehat{\lambda}_0 = \mathbb{1}[X_1 = 1] \cdot$ (correction) via conditioning on S_N. In particular, the UMVUE of $e^{-\lambda} = \Pr(X = 0)$ is $(1 - 1/N)^{S_N}$ — the fraction of the data that is zero, raised to the N-th power, a classic result derivable in two lines from the Lehmann-Scheffé theorem but requiring careful moment calculation otherwise.

Moments of the Sufficient Statistic via the Log-Partition Function

The log-partition function $\Phi(\theta)$ is not merely a normalising constant: its derivatives generate the cumulants of $T(X)$, exactly analogous to a moment/cumulant generating function in probability (Volume I, Part 1).

For $p_\theta(x) = h(x)\exp(\theta^\top T(x) - \Phi(\theta))$,

$$\nabla\Phi(\theta) = \mathbb{E}_{p_\theta}[T(X)] \triangleq \eta(\theta), \qquad \nabla^2\Phi(\theta) = \mathrm{Cov}_{p_\theta}(T(X)).$$

Proof. From $\Phi(\theta) = \log \int h(x)e^{\theta^\top T(x)}dx$,

$$\partial_i\Phi(\theta) = \frac{\int h(x)T_i(x)e^{\theta^\top T(x)}dx}{\int h(x)e^{\theta^\top T(x)}dx} = \int T_i(x)\, p_\theta(x)\, dx = \mathbb{E}_{p_\theta}[T_i(X)].$$

Differentiating again, $\partial_i\partial_j\Phi(\theta) = \partial_j\mathbb{E}_{p_\theta}[T_i(X)] = \int T_i(x)\,\partial_j p_\theta(x)\,dx = \int T_i(x)(T_j(x) - \partial_j\Phi(\theta))p_\theta(x)\,dx = \mathbb{E}[T_iT_j] - \mathbb{E}[T_i]\mathbb{E}[T_j] = \mathrm{Cov}(T_i, T_j)$, using $\partial_j p_\theta = (T_j(x) - \partial_j\Phi(\theta))p_\theta(x)$ (differentiate the canonical form directly). ■

For an exponential family in natural parameters θ,

$$\mathbf{G}(\theta) = \nabla^2\Phi(\theta) = \mathrm{Cov}_{p_\theta}(T(X)).$$

Proof. $\log p_\theta(x) = \log h(x) + \theta^\top T(x) - \Phi(\theta)$, so $\partial_i\partial_j \log p_\theta(x) = -\partial_i\partial_j\Phi(\theta)$ (the only θ-dependence in $\log p_\theta$ besides the linear term is $-\Phi(\theta)$, and the linear term $\theta^\top T(x)$ has zero second derivative in θ). By Theorem 1.2.2, $g_{ij}(\theta) = -\mathbb{E}[\partial_i\partial_j \log p_\theta] = \partial_i\partial_j\Phi(\theta)$. Combined with Theorem 2.1.4, this equals $\mathrm{Cov}(T_i, T_j)$. ■

This single identity, $\mathbf{G}(\theta) = \nabla^2\Phi(\theta)$, is the seed of everything in Parts 2 and 3: *the Fisher information matrix is the Hessian of a single convex scalar function* Φ. A Riemannian metric that is globally the Hessian of a convex potential is called a *Hessian metric*, and Hessian metrics have an automatic dual-coordinate structure via the Legendre transform of the potential — which is exactly the θ/η duality of Section 2.3. Convexity of Φ (equivalent to $\mathbf{G} \succeq 0$, which holds always since $\mathbf{G}$ is a covariance matrix) is therefore not an extra assumption but a restatement of positive-semidefiniteness of the Fisher metric, now seen as a consequence of Φ being a cumulant generating function.

Summary of Section 2.1. Every exponential family can be written in canonical form $p_\theta(x) = h(x)\exp(\theta^\top T(x) - \Phi(\theta))$, where θ is the natural parameter and $T(x)$ is the sufficient statistic. The Pitman–Koopman–Darmois theorem characterises exponential families as exactly those regular models with fixed-dimensional sufficient statistics. The log-partition function $\Phi(\theta)$ is convex, with $\nabla\Phi(\theta) = \mathbb{E}_\theta[T(X)] \triangleq \eta(\theta)$ and $\nabla^2\Phi(\theta) = \mathrm{Cov}_\theta(T(X)) = \mathbf{G}(\theta)$: the Fisher metric is the Hessian of Φ, a Hessian metric — the structural fact from which the dual coordinate geometry of the rest of Part 2 follows.

2.2 The Cumulant Generating Function and Legendre Duality

Convex Conjugates: a Brief Recap

Legendre–Fenchel Conjugate

For a convex function $\Phi : \Theta \to \mathbb{R}$, its *convex conjugate* (Legendre–Fenchel transform) is

$$\Psi(\eta) \triangleq \sup_{\theta\in\Theta} \left[\theta^\top\eta - \Phi(\theta)\right].$$

If Φ is strictly convex and differentiable, the supremum is attained at the unique θ solving $\nabla\Phi(\theta) = \eta$, and Ψ is itself strictly convex and differentiable on its domain, with $\nabla\Psi(\eta) = \theta$, i.e. $\nabla\Psi = (\nabla\Phi)^{-1}$. The pair (Φ, Ψ) are *Legendre duals* of each other: $(\Psi)^* = \Phi$ as well (biduality, for closed convex functions).

For an exponential family, $\theta \mapsto \eta = \nabla\Phi(\theta) = \mathbb{E}_\theta[T(X)]$ (Theorem 2.1.4) is exactly the gradient map whose inverse defines Ψ. The variable η thus has a direct statistical meaning: it is the *expectation parameter* — the expected value of the sufficient statistic.

Expectation Parameters

The map $\theta \mapsto \eta \triangleq \mathbb{E}_\theta[T(X)] = \nabla\Phi(\theta)$ is called the *θ-to-η coordinate change*, and η are the *expectation* (or *moment*, or *dual*) *coordinates* on $\mathcal{S}$. By the inverse function theorem (since $\nabla^2\Phi = \mathbf{G} \succ 0$ under identifiability), this map is a diffeomorphism between the natural parameter space Θ and the expectation parameter space $\mathcal{H} = \nabla\Phi(\Theta)$.

The Dual Potential is (Minus) the Entropy

For an exponential family, $\Psi(\eta) = \theta(\eta)^\top\eta - \Phi(\theta(\eta)) = -H(p_{\theta(\eta)}) + \mathbb{E}_{p_{\theta(\eta)}}[\log h(X)]$, where $H(p) = -\mathbb{E}_p[\log p(X)]$ is the Shannon entropy (Volume I). When $h(x) \equiv \text{const}$ (e.g. Bernoulli, categorical, exponential), $\Psi(\eta) = -H(p_{\theta(\eta)}) + \text{const}$: *the convex conjugate of the log-partition function is, up to an additive constant, the negative Shannon entropy, expressed as a function of the mean parameters.*

Proof. $\log p_\theta(x) = \log h(x) + \theta^\top T(x) - \Phi(\theta)$, so $-H(p_\theta) = \mathbb{E}_\theta[\log p_\theta(X)] = \mathbb{E}_\theta[\log h(X)] + \theta^\top\mathbb{E}_\theta[T(X)] - \Phi(\theta) = \mathbb{E}_\theta[\log h(X)] + \theta^\top\eta - \Phi(\theta)$. By definition of the conjugate, $\theta^\top\eta - \Phi(\theta) = \Psi(\eta)$ at the value $\theta = \theta(\eta)$ maximising $\theta^\top\eta - \Phi(\theta)$ (which holds at $\theta = \theta(\eta)$ since $\nabla\Phi(\theta(\eta)) = \eta$ by definition of $\theta(\eta)$ as the inverse of the gradient map). Hence $-H(p_{\theta(\eta)}) = \mathbb{E}[\log h] + \Psi(\eta)$, i.e. $\Psi(\eta) = -H(p_{\theta(\eta)}) - \mathbb{E}[\log h(X)]$. ■

This theorem is the reason Ψ is traditionally called the *negentropy* function in information geometry: **the natural parameters θ are dual, via the Legendre transform of the entropy, to the expectation parameters η, and the two convex potentials Φ (log-partition / free energy) and Ψ (negative entropy) are Legendre conjugates of each other.** This is the information-geometric incarnation of the thermodynamic free-energy/entropy duality, and it is no accident: the log-partition function $\Phi(\theta)$ in statistical mechanics literally *is* the (negative) free energy divided by temperature, with θ playing the role of inverse temperature times the conjugate thermodynamic variable, and $\eta = \nabla\Phi(\theta)$ the corresponding extensive variable (e.g. energy).

Worked Dual Potentials

Example — Bernoulli: Φ and Ψ

With natural parameter $\theta = \log\frac{p}{1-p}$ (logit) and $\Phi(\theta) = \log(1+e^\theta)$, the expectation parameter is $\eta = \nabla\Phi(\theta) = \frac{e^\theta}{1+e^\theta} = p$ — the success probability itself. The dual potential is

$$\Psi(\eta) = \theta\eta - \Phi(\theta)\big|_{\theta=\text{logit}(\eta)} = \eta\log\eta + (1-\eta)\log(1-\eta),$$

which is exactly $-H(\text{Bernoulli}(\eta))$ (binary entropy with a minus sign), confirming Theorem 2.2.2 with $h \equiv 1$. The Fisher metric in η-coordinates is $\nabla^2_\eta\Psi(\eta) = \frac{1}{\eta(1-\eta)} = \mathbf{G}(\eta)^{-1}$ – wait, more precisely: $\nabla^2\Psi(\eta) = [\nabla^2\Phi(\theta)]^{-1}$ evaluated at the corresponding θ (Theorem 2.2.3 below), which here gives $1/(\eta(1-\eta))$, matching $g(\theta)$ from Example 1.5.1 expressed in the $\eta = p$ coordinate (since the Fisher metric in η-coordinates is the *inverse* of $\mathbf{G}$ in θ-coordinates evaluated at the dual point — a relation made precise next).

Example — Univariate Gaussian: Φ, η, and Ψ

With $T(x) = (x, x^2)$ and natural parameters $\theta = (\theta_1, \theta_2) = (\mu/\sigma^2, -1/(2\sigma^2))$ (Example 2.1.2), the log-partition function is

$$\Phi(\theta) = -\frac{\theta_1^2}{4\theta_2} - \tfrac{1}{2}\log(-2\theta_2) + \tfrac{1}{2}\log\pi.$$

The expectation parameters are $\eta_1 = \mathbb{E}[X] = \mu$ and $\eta_2 = \mathbb{E}[X^2] = \mu^2 + \sigma^2$, so the mean-parameter coordinates $(\eta_1, \eta_2) = (\mu, \mu^2+\sigma^2)$ are *first and second moments*, not (μ, σ^2) directly — the η-coordinates "mix" mean and variance in a way that is initially counter-intuitive but is exactly what makes them the Legendre-dual coordinates. The dual potential, after substitution, works out to

$$\Psi(\eta) = -\tfrac{1}{2}\log(\eta_2 - \eta_1^2) + \text{const} = -\tfrac{1}{2}\log\sigma^2 + \text{const},$$

matching $-H(\mathcal{N}(\mu,\sigma^2)) = -\frac{1}{2}\log(2\pi e\sigma^2)$ up to the additive constant, exactly as Theorem 2.2.2 predicts.

Example — Poisson: Φ, η, and Ψ

For Poi(λ) with sufficient statistic $T(x) = x$, the canonical form is $p_\theta(x) = \exp(\theta x - e^\theta)/x!$ with natural parameter $\theta = \log\lambda$ and log-partition $\Phi(\theta) = e^\theta$. The expectation parameter is $\eta = \nabla\Phi(\theta) = e^\theta = \lambda = \mathbb{E}[X]$ (the Poisson mean), so the η-coordinate is

simply the rate λ. The dual potential is:

$$\Psi(\eta) = \theta\eta - \Phi(\theta)\big|_{\theta=\log\eta} = \eta\log\eta - \eta = \lambda\log\lambda - \lambda,$$

and indeed $-H(\mathrm{Poi}(\lambda)) = -\lambda\log\lambda + \lambda - \frac{1}{2}\log(2\pi e\lambda) + O(1/\lambda)$ (Stirling), so $\Psi(\eta)$ matches the leading terms. The inverse-Hessian relation: $\nabla^2\Phi(\theta) = e^\theta = \lambda = \eta$ (Fisher information in θ-coordinate), $\nabla^2\Psi(\eta) = 1/\eta = 1/\lambda$ (its inverse in η-coordinate), confirming $\nabla^2\Psi = [\nabla^2\Phi]^{-1}$. The KL divergence in dual coordinates:

$$D_{\mathrm{KL}}(\mathrm{Poi}(\lambda')\,\|\,\mathrm{Poi}(\lambda)) = \lambda' - \lambda + \lambda'\log(\lambda/\lambda') = B_\Psi(\eta' : \eta) = \Psi(\eta') - \Psi(\eta) - \nabla\Psi(\eta)^\top(\eta' - \eta),$$

the Bregman divergence of $\Psi(\eta) = \eta\log\eta - \eta$ (the "negative Poisson entropy potential"), as expected from Theorem 2.3.4.

Example — Gamma: Φ, η, and Ψ

For $\Gamma(\alpha, \beta)$ with density $p(x;\alpha,\beta) = \beta^\alpha x^{\alpha-1}e^{-\beta x}/\Gamma(\alpha)$, the sufficient statistics are $T(x) = (\log x, x)$ and the natural parameters are $\theta = (\alpha - 1, -\beta)$. The log-partition is $\Phi(\theta_1, \theta_2) = \log\Gamma(\theta_1 + 1) - (\theta_1 + 1)\log(-\theta_2)$. The expectation parameters are:

$$\eta_1 = \mathbb{E}[\log X] = \psi(\alpha) - \log\beta, \qquad \eta_2 = \mathbb{E}[X] = \alpha/\beta.$$

The dual potential $\Psi(\eta_1, \eta_2)$ is the negative entropy $-H(\Gamma(\alpha,\beta))$ where (α, β) are expressed in terms of (η_1, η_2) via the system above. No closed form exists for the inverse (it requires solving $\psi(\alpha) = \eta_1 + \log(\alpha/\eta_2)$ numerically), but the Legendre structure guarantees the conjugate exists and is strictly convex. This illustrates the general principle that the η-parametrisation is "natural" for Bayesian updating (the posterior sufficient statistic is $\bar{T}_n = \sum_i T(x_i)/n$, and the MLE is the η matching $\bar{T}_n$) while the θ-parametrisation is "natural" for exponential family algebra, and neither is easy to invert analytically for the Gamma family.

The Inverse Hessian Relation

At corresponding points $\eta = \nabla\Phi(\theta)$ (equivalently $\theta = \nabla\Psi(\eta)$), the Hessians of the dual potentials are matrix inverses of each other:

$$\nabla^2_\eta\Psi(\eta) = \left[\nabla^2_\theta\Phi(\theta)\right]^{-1} = \mathbf{G}(\theta)^{-1}.$$

Proof. Differentiate the identity $\eta = \nabla\Phi(\theta(\eta))$ with respect to η (where $\theta(\eta) = \nabla\Psi(\eta)$ is the inverse coordinate map): by the chain rule, $\mathbf{I} = \nabla^2\Phi(\theta)\cdot\nabla\theta(\eta) = \nabla^2\Phi(\theta)\cdot\nabla^2\Psi(\eta)$ (since $\nabla\theta(\eta) = \nabla\nabla\Psi(\eta) = \nabla^2\Psi(\eta)$). Hence $\nabla^2\Psi(\eta) = [\nabla^2\Phi(\theta)]^{-1}$. ■

This identity is the precise statement that θ- and η-coordinates carry *dual* metrics: if $\mathbf{G}(\theta) = \nabla^2\Phi(\theta)$ is the Fisher metric in natural coordinates, then $\mathbf{G}(\theta)^{-1} = \nabla^2\Psi(\eta)$ is (by the same Hessian-metric construction applied to Ψ) the Fisher metric *in η-coordinates*, at the corresponding point. Both describe the same Riemannian manifold $(\mathcal{S}, \mathbf{G})$, just in two different coordinate systems related by the Legendre transform — and the inverse-matrix relation between the two Hessians is the local shadow of this global duality. In Part 3, this duality is upgraded from a statement about metrics to a statement about *connections*: the θ-coordinates

are affine for the e-connection, the η-coordinates are affine for the m-connection, and these are dual connections with respect to $\mathbf{G}$ in a precise sense that generalises Theorem 2.2.3.

Summary of Section 2.2. The expectation parameters $\eta = \nabla\Phi(\theta) = \mathbb{E}_\theta[T(X)]$ are Legendre-dual to the natural parameters θ, via the conjugate potential $\Psi(\eta) = \sup_\theta[\theta^\top\eta - \Phi(\theta)]$, which equals (minus) the Shannon entropy as a function of the mean parameters. The Hessians of the two dual potentials are matrix inverses at corresponding points, $\nabla^2\Psi(\eta) = \mathbf{G}(\theta)^{-1}$, expressing the Fisher metric in η-coordinates as the inverse of the Fisher metric in θ-coordinates. This $\theta \leftrightarrow \eta$ duality, generated by the pair (Φ, Ψ), is the coordinate-level manifestation of the dual affine-connection structure of Part 3.

2.3 Bregman Divergences and KL Divergence as the Canonical Example

Bregman Divergences: Definition and First Properties

Sections 2.1–2.2 established that, for an exponential family, the Fisher metric is the Hessian of a convex potential Φ, and that the Legendre dual Ψ of Φ is (minus) the entropy in expectation coordinates. This section introduces the divergence functional naturally associated with any convex potential — the *Bregman divergence* — and proves the central theorem of Part 2: the KL divergence between two members of an exponential family is *exactly* a Bregman divergence between their parameters.

Bregman Divergence

Let $\phi : \Omega \to \mathbb{R}$ be a strictly convex, differentiable function on a convex domain $\Omega \subseteq \mathbb{R}^n$. The *Bregman divergence* generated by ϕ is

$$B_\phi(p : q) \triangleq \phi(p) - \phi(q) - \nabla\phi(q)^\top(p - q), \qquad p, q \in \Omega.$$

Geometrically, $\phi(q) + \nabla\phi(q)^\top(p-q)$ is the value, at p, of the first-order Taylor expansion (tangent hyperplane) of ϕ at q. The Bregman divergence is therefore the *vertical gap* between $\phi(p)$ and the tangent-plane approximation to ϕ at q, evaluated at p — see Figure-style intuition in the Numerical Tour below.

For ϕ strictly convex and differentiable on Ω:

1. **Non-negativity.** $B_\phi(p : q) \geq 0$ for all $p, q \in \Omega$, with equality iff $p = q$.
2. **Convexity in the first argument.** $p \mapsto B_\phi(p : q)$ is strictly convex (it differs from $\phi(p)$ by an affine function of p).
3. **Asymmetry.** In general $B_\phi(p : q) \neq B_\phi(q : p)$, and the triangle inequality fails: a Bregman divergence is *not* a metric.
4. **Three-point identity.** For all $p, q, r \in \Omega$,
$$B_\phi(p : q) = B_\phi(p : r) + B_\phi(r : q) + (\nabla\phi(r) - \nabla\phi(q))^\top(p - r).$$
5. **Invariance under affine terms.** If $\phi'(x) = \phi(x) + a^\top x + b$ for a fixed vector a and scalar b, then $B_{\phi'}(p : q) = B_\phi(p : q)$: adding a linear function to ϕ does not change the divergence (the linear term's contribution to $\phi(p) - \phi(q)$ is exactly cancelled by its contribution to $\nabla\phi(q)^\top(p - q)$).

Proof of (1). By strict convexity, the first-order Taylor remainder $\phi(p) - \phi(q) - \nabla\phi(q)^\top(p - q)$ equals $\frac{1}{2}(p-q)^\top\nabla^2\phi(\xi)(p-q)$ for some ξ on the segment $[p, q]$ (Taylor's theorem with Lagrange remainder), and $\nabla^2\phi(\xi) \succ 0$ by strict convexity, so this quadratic form is > 0 unless $p = q$. ■

Proof of (4). Expand the right-hand side: $B_\phi(p : r) + B_\phi(r : q) = [\phi(p) - \phi(r) - \nabla\phi(r)^\top(p - r)] + [\phi(r) - \phi(q) - \nabla\phi(q)^\top(r - q)]$. The $\phi(r)$ terms cancel: $= \phi(p) - \phi(q) - \nabla\phi(r)^\top(p - r) - \nabla\phi(q)^\top(r - q)$. Add and subtract $\nabla\phi(q)^\top(p - r)$: $= \phi(p) - \phi(q) - \nabla\phi(q)^\top(p - q) + (\nabla\phi(q) - \nabla\phi(r))^\top(p - r) = B_\phi(p : q) + (\nabla\phi(q) - \nabla\phi(r))^\top(p - r)$. Rearranging gives the stated identity (with a sign flip on the last term, matching the displayed form). ■

Example — The Euclidean Case: $\phi(x) = \|x\|^2$

For $\phi(x) = \|x\|^2 = x^\top x$, $\nabla\phi(x) = 2x$, so

$$B_\phi(p : q) = \|p\|^2 - \|q\|^2 - 2q^\top(p - q) = \|p\|^2 - 2p^\top q + \|q\|^2 = \|p - q\|^2 .$$

The Bregman divergence generated by the squared Euclidean norm is the squared Euclidean distance. In this case $B_\phi(p : q) = B_\phi(q : p)$ (the rare symmetric case — symmetry of a Bregman divergence forces ϕ to be quadratic), and the three-point identity of Property 4 reduces to the algebraic identity $\|p - q\|^2 = \|p - r\|^2 + \|r - q\|^2 + 2(r - q)^\top(p - r)$, which is just the law of cosines / parallelogram identity in disguise.

The Main Theorem: KL Divergence as a Bregman Divergence of the Log-Partition Function

Let $\{p_\theta : \theta \in \Theta\}$ be an exponential family in canonical form, $p_\theta(x) = h(x)\exp(\theta^\top T(x) - \Phi(\theta))$. For any $\theta, \theta' \in \Theta$,

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta) .$$

Proof. By definition,

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = \mathbb{E}_{p_\theta}\left[\log \frac{p_\theta(X)}{p_{\theta'}(X)}\right] = \mathbb{E}_{p_\theta}\left[(\theta - \theta')^\top T(X) - \Phi(\theta) + \Phi(\theta')\right] ,$$

since the carrier $h(x)$ appears in both numerator and denominator and cancels, leaving only the exponent difference $\theta^\top T(x) - \Phi(\theta) - (\theta'^\top T(x) - \Phi(\theta')) = (\theta - \theta')^\top T(x) - \Phi(\theta) + \Phi(\theta')$. Taking the expectation under p_θ and using $\mathbb{E}_{p_\theta}[T(X)] = \nabla\Phi(\theta)$ (Theorem 2.1.4),

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = (\theta - \theta')^\top\nabla\Phi(\theta) - \Phi(\theta) + \Phi(\theta') = \Phi(\theta') - \Phi(\theta) - \nabla\Phi(\theta)^\top(\theta' - \theta) = B_\Phi(\theta' : \theta) ,$$

matching Definition 2.3.1 with $\phi = \Phi$, evaluated at the pair (θ', θ). ■

The argument-order reversal. Note carefully: $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'})$ — KL "from θ to θ'", i.e. θ is the reference (first) distribution — equals $B_\Phi(\theta' : \theta)$, the Bregman divergence "from θ' to θ". The two orderings are *swapped*. This is not a typo: it follows necessarily from the asymmetry of both objects and the direction of the Taylor expansion (the Bregman divergence is the gap above the tangent plane taken *at the second argu-*

ment, while the KL expectation is taken *under the first*). Tracking this swap carefully is essential when applying the theorem — a sign or argument-order error here is the single most common mistake when working with information-geometric divergence identities.

Example — Verifying the Identity for Bernoulli

Let $\theta = \text{logit}(p)$, $\theta' = \text{logit}(p')$, $\Phi(\theta) = \log(1+e^\theta)$ (Example 2.1.1). Direct computation: $\nabla\Phi(\theta) = \frac{e^\theta}{1+e^\theta} = p$, so

$$B_\Phi(\theta' : \theta) = \Phi(\theta') - \Phi(\theta) - p(\theta' - \theta) = \log\frac{1+e^{\theta'}}{1+e^\theta} - p(\theta'-\theta).$$

Take $p = 0.3$ ($\theta = \text{logit}(0.3) \approx -0.8473$) and $p' = 0.6$ ($\theta' \approx 0.4055$). Then $\Phi(\theta) = \log(1+e^{-0.8473}) = \log(1.4286) \approx 0.3567$, $\Phi(\theta') = \log(1+e^{0.4055}) = \log(2.5001) \approx 0.9163$, and $\theta' - \theta \approx 1.2528$. So $B_\Phi(\theta' : \theta) \approx 0.9163 - 0.3567 - 0.3 \times 1.2528 \approx 0.5596 - 0.3758 \approx 0.1838$ nats.
Directly: $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = D_{\mathrm{KL}}(\text{Bern}(0.3) \,\|\, \text{Bern}(0.6)) = 0.3\ln\frac{0.3}{0.6} + 0.7\ln\frac{0.7}{0.4} = 0.3 \times (-0.6931) + 0.7 \times 0.5596 \approx -0.2079 + 0.3917 \approx 0.1838$ nats. The two computations agree to four decimal places, confirming Theorem 2.3.2.

The Dual Form: KL as a Bregman Divergence of the Negentropy

The Legendre-duality machinery of Section 2.2 yields a second, dual expression for the same KL divergence, this time in expectation (η) coordinates.

Let ϕ and ϕ^* be Legendre-dual convex functions, with $\eta = \nabla\phi(\theta) \Leftrightarrow \theta = \nabla\phi^*(\eta)$. Then for any θ, θ', writing $\eta = \nabla\phi(\theta)$, $\eta' = \nabla\phi(\theta')$,

$$B_\phi(\theta' : \theta) = B_{\phi^*}(\eta : \eta').$$

Proof sketch. By definition of the Legendre transform, $\phi(\theta) = \theta^\top\eta - \phi^*(\eta)$ and $\phi(\theta') = \theta'^\top\eta' - \phi^*(\eta')$ (each θ paired with its own dual η). Substitute into $B_\phi(\theta' : \theta) = \phi(\theta') - \phi(\theta) - \eta^\top(\theta' - \theta)$ (using $\nabla\phi(\theta) = \eta$):

$$= [\theta'^\top\eta' - \phi^*(\eta')] - [\theta^\top\eta - \phi^*(\eta)] - \eta^\top\theta' + \eta^\top\theta = \theta'^\top\eta' - \theta'^\top\eta - [\phi^*(\eta') - \phi^*(\eta)] = \theta'^\top(\eta' - \eta) - [\phi^*(\eta') - \phi^*(\eta)].$$

Since $\theta' = \nabla\phi^*(\eta')$, this is $\nabla\phi^*(\eta')^\top(\eta' - \eta) - \phi^*(\eta') + \phi^*(\eta) = -[\phi^*(\eta) - \phi^*(\eta') - \nabla\phi^*(\eta')^\top(\eta - \eta')]$ ... a short rearrangement (left as an exercise in index bookkeeping) shows this equals $B_{\phi^*}(\eta : \eta') = \phi^*(\eta) - \phi^*(\eta') - \nabla\phi^*(\eta')^\top(\eta - \eta')$. ■

Applying this with $\phi = \Phi$, $\phi^* = \Psi$, $\eta = \nabla\Phi(\theta)$, $\eta' = \nabla\Phi(\theta')$:

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta) = B_\Psi(\eta : \eta').$$

Dual expressions for KL divergence.

$$\boxed{D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta) = B_\Psi(\eta : \eta')}$$

The natural-coordinate Bregman divergence (generated by Φ, evaluated with arguments *reversed*) and the expectation-coordinate Bregman divergence (generated by Ψ, arguments *not* reversed) compute the *same number*. Both formulas are useful: the Ψ-form is often the more familiar one, since $\Psi(\eta) = -H(p_\eta) + \text{const}$ (Theorem 2.2.2) makes $B_\Psi(\eta : \eta')$ a difference-of-entropies-plus-gradient expression that, for many families, reduces directly to the textbook KL formula (see Section 2.5).

Why This Matters: Divergence as Geometry, Not Just Algebra

The identity $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta)$ recasts the KL divergence — defined in Volume I purely as an expectation of a log-likelihood ratio — as a *geometric* quantity on the manifold $\mathcal{S}$: the gap between Φ and its tangent plane. This has three immediate consequences, each developed further later in the volume.

- **Local quadratic behaviour.** For $\theta' = \theta + d\theta$ infinitesimally close to θ, Taylor expansion of $B_\Phi(\theta' : \theta)$ to second order gives $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+d\theta}) \approx \frac{1}{2} d\theta^\top \nabla^2 \Phi(\theta)\, d\theta = \frac{1}{2} d\theta^\top \mathbf{G}(\theta)\, d\theta$ — recovering the second-order/Fisher-metric relation between KL and the Riemannian distance already previewed in Section 1.2 (Formula Reference, Part 1: "$\partial_{\theta'^i}\partial_{\theta'^j} D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'})\big|_{\theta'=\theta} = g_{ij}(\theta)$"). The Bregman-divergence viewpoint makes this exact second-order relation a special case of a much more general (all-orders, not just infinitesimal) statement.
- **Asymmetry is geometric, not pathological.** KL divergence's notorious asymmetry ($D_{\mathrm{KL}}(p \,\|\, q) \neq D_{\mathrm{KL}}(q \,\|\, p)$) is, via Theorem 2.3.2, simply the asymmetry of the Bregman divergence (Property 3) — itself a direct, unavoidable consequence of Φ being a *nonlinear* convex function (linear Φ would give a symmetric, trivial divergence). Asymmetry is therefore not a defect to be corrected, but the price of $\mathcal{S}$ having genuinely curved (non-Euclidean) geometry.
- **A geometric route to Pythagorean-type theorems.** Because $B_\phi(\cdot : \cdot)$ behaves, in many respects, like a squared distance (Property 1: non-negative, zero only on the diagonal; Example 2.3.3: literally a squared distance for $\phi = \|\cdot\|^2$), one can ask whether Bregman divergences satisfy a Pythagorean theorem for projections onto convex sets, generalising the Euclidean orthogonal-projection Pythagorean theorem. Section 2.4 shows that they do — and that, via Theorem 2.3.2, this immediately yields a Pythagorean theorem for KL divergence itself, the foundation of information projections (Part 4).

Summary of Section 2.3. The Bregman divergence $B_\phi(p : q) = \phi(p) - \phi(q) - \nabla\phi(q)^\top(p - q)$, generated by any strictly convex ϕ, is non-negative, vanishes only on the diagonal, and satisfies a three-point identity (Property 4) that generalises the law of cosines. For exponential families, $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta) = B_\Psi(\eta : \eta')$ (Theorem 2.3.2 and its dual): KL divergence *is* a Bregman divergence, in either of two dual coordinate systems, with an argument-order reversal in the θ-form. This single identity is the algebraic core of information geometry's contribution to statistics: every Bregman-divergence theorem (Pythagorean decompositions, projection algorithms, convergence of alternating minimisation) becomes, via Theorem 2.3.2, a theorem about KL divergence and exponential families.

2.4 The Generalized Pythagorean Theorem for Bregman Divergences

Bregman Projections

Bregman Projection

Let ϕ be a strictly convex, differentiable function on $\Omega \subseteq \mathbb{R}^n$, and let $C \subseteq \Omega$ be closed and convex. The *Bregman projection* of $p \in \Omega$ onto C is

$$\Pi_C^\phi(p) \triangleq \arg\min_{r \in C} B_\phi(p : r).$$

Because $r \mapsto B_\phi(p : r)$ is, in general, *not* convex in r (only the first argument of a Bregman divergence is guaranteed convex, by Property 2 of Section 2.3), existence and uniqueness of $\Pi_C^\phi(p)$ require some care; we assume throughout that ϕ is such that $B_\phi(p : \cdot)$ is, in fact, strictly convex on C (this holds for all the families used in this volume, where $\phi \in \{\Phi, \Psi\}$ and C is itself an affine slice of Θ or of the η-space).

Let ϕ be strictly convex and $C \subseteq \operatorname{dom}\phi$ closed and convex. Suppose $r \mapsto B_\phi(p : r)$ is strictly convex on C and coercive (tends to $+\infty$ as $r \to \partial C$ or $\|r\| \to \infty$). Then:

1. **Existence and uniqueness.** $\Pi_C^\phi(p)$ exists and is unique.
2. **First-order characterisation.** $q^* = \Pi_C^\phi(p)$ if and only if, for all $r \in C$,

$$(\nabla\phi(p) - \nabla\phi(q^*))^\top (r - q^*) \le 0.$$

   For C affine this becomes an equality (the gradient $\nabla\phi(p) - \nabla\phi(q^*)$ is g-orthogonal to C at q^*).
3. **Non-expansiveness.** For any $r \in C$, $B_\phi\big(r : \Pi_C^\phi(p)\big) \le B_\phi(r : p)$: the projection moves every point *closer* (in the first-argument Bregman sense) to any point in C. In particular, successive projections Π_C^ϕ converge.

Alternating Bregman projection algorithm. When the constraint set is the intersection $C = C_1 \cap C_2$ of two closed convex sets C_1, C_2 (both affine, for simplicity), the *alternating Bregman projection* algorithm generates the sequence:

$$p^{(2t-1)} = \Pi_{C_1}^\phi(p^{(2t-2)}), \qquad p^{(2t)} = \Pi_{C_2}^\phi(p^{(2t-1)}), \quad t = 1, 2, \ldots$$

starting from any $p^{(0)} \in \operatorname{dom}\phi$. Under mild regularity conditions (Csiszár and Tusnády 1984), this sequence converges to $\Pi_{C_1 \cap C_2}^\phi(p^{(0)})$. The Pythagorean theorem (Theorem 2.4.2 below) gives, at each step,

$$B_\phi\big(p^* : p^{(t+1)}\big) \le B_\phi\big(p^* : p^{(t)}\big) - B_\phi\big(p^{(t+1)} : p^{(t)}\big),$$

so the distance to the solution decreases by the amount of "Bregman slack" $B_\phi\big(p^{(t+1)} : p^{(t)}\big)$ at each step. This is the abstract form of the EM monotone-improvement property (Section 4.2) and of the IPF convergence proof (Section 4.3).

Example — Alternating Projections for a 2×2 Joint Distribution

Let $\phi = -H$ (negative entropy), so the Bregman projection onto a marginal constraint is an IPF step. Take the reference $q = \mathrm{Unif}(\{00, 01, 10, 11\}) = (1/4, 1/4, 1/4, 1/4)$ and

constraint sets $C_1 = \{p : p_{X_1} = (3/4, 1/4)\}$ and $C_2 = \{p : p_{X_2} = (1/2, 1/2)\}$.
Step 1 ($\Pi^{\phi}_{C_1}$): rescale rows so first marginal is $(3/4, 1/4)$: $p^{(1)} = (3/8, 3/8, 1/8, 1/8)$.
Step 2 ($\Pi^{\phi}_{C_2}$): rescale columns so second marginal is $(1/2, 1/2)$: $p^{(2)} = (3/8, 3/8, 1/8, 1/8) \cdot (1/2, 1/2)/(3/4, 1/4)$ (columnwise), giving $p^{(2)} = (3/8 \cdot 2/3, 3/8 \cdot 2/1, \ldots)$ — computed as: column 0 target $1/2$, current sum $3/8 + 1/8 = 1/2$, so column 0 unchanged; column 1 target $1/2$, current $1/2$, also unchanged. Since both marginals were already achieved, $p^{(2)} = p^{(1)} = p^*$. The intersection $C_1 \cap C_2$ has a unique element (given independence under uniform reference, verified by $p^*_{XY} = p^*_X \cdot p^*_Y$ in this case), and the algorithm converges in one step because the two constraint submanifolds are g-orthogonal at p^*.

The analogy between alternating Bregman projections and the method of alternating orthogonal projections in Hilbert spaces (von Neumann's theorem, 1950: alternating orthogonal projections onto two closed subspaces converge to the projection onto the intersection) is not an analogy but a special case: for $\phi(x) = \frac{1}{2}\|x\|^2$, all Bregman projections are Euclidean orthogonal projections, and the convergence theorem reduces to von Neumann's. The information-geometric generalisation replaces the Hilbert-space inner product by the Fisher metric and the linear projections by the dual geodesic projections of Section 3.3, with the Pythagorean theorem as the central tool in both settings.

The Generalized Pythagorean Theorem

Let $C \subseteq \Omega$ be convex, $p \in \Omega$, and $q^* = \Pi^{\phi}_C(p)$. Then for every $r \in C$,

$$B_\phi(p : r) \;\geq\; B_\phi(p : q^*) + B_\phi(q^* : r)\,.$$

If, in addition, C is an *affine* subset of Ω (i.e. $C = \{x : Ax = b\}$ for some matrix A and vector b), then equality holds for every $r \in C$:

$$B_\phi(p : r) = B_\phi(p : q^*) + B_\phi(q^* : r)\,, \qquad \forall r \in C.$$

Proof. Apply the three-point identity (Property 4 of Section 2.3) with the triple (p, q^*, r):

$$B_\phi(p : r) = B_\phi(p : q^*) + B_\phi(q^* : r) + (\nabla\phi(q^*) - \nabla\phi(r))^\top (p - q^*).$$

It remains to show the cross term $(\nabla\phi(q^*) - \nabla\phi(r))^\top (p - q^*)$ is ≥ 0 for general convex C, and $= 0$ for affine C.

General convex C. Since q^* minimises $r \mapsto B_\phi(p : r)$ over C, the first-order optimality (variational inequality) condition states that, for every $r \in C$,

$$\nabla_r B_\phi(p : r)\Big|^{\top}_{r=q^*} (r - q^*) \;\geq\; 0.$$

Computing $\nabla_r B_\phi(p : r) = -\nabla^2\phi(r)(p - r)$ (differentiate $\phi(p) - \phi(r) - \nabla\phi(r)^\top(p - r)$ with respect to r, using the product rule on the last term), the optimality condition reads $-\nabla^2\phi(q^*)(p - q^*)^\top (r - q^*) \geq 0$, i.e. $(p - q^*)^\top \nabla^2\phi(q^*)(r - q^*) \leq 0$. This is *not* immediately the cross term we want (which involves $\nabla\phi(q^*) - \nabla\phi(r)$, not $\nabla^2\phi(q^*)$); a short additional argument — using the mean-value form $\nabla\phi(r) - \nabla\phi(q^*) = \nabla^2\phi(\xi)(r - q^*)$ for some

$\xi \in [q^*, r]$ and a convexity comparison between $\nabla^2\phi(\xi)$ and $\nabla^2\phi(q^*)$ — establishes the inequality $(\nabla\phi(q^*) - \nabla\phi(r))^\top(p - q^*) \geq 0$ in general; we omit the technical details and refer to Csiszár (1975) for the fully general statement (which holds even for ϕ merely convex, not necessarily C^2, via subgradients).

Affine C. If $C = \{x : Ax = b\}$, then $q^* \in C$ and $r \in C$ imply $A(r - q^*) = 0$, i.e. $r - q^*$ lies in the null space of A. The first-order optimality condition for minimising $B_\phi(p : \cdot)$ over the affine set C is that $\nabla_r B_\phi(p : r)\,|_{r=q^*}$ is orthogonal to the null space of A (equivalently, $\nabla_r B_\phi(p : r)\,|_{r=q^*} \in \mathrm{row}(A)$, by Lagrangian stationarity with multiplier λ: $\nabla^2\phi(q^*)(p - q^*) = A^\top\lambda$ for some λ). Now the cross term is $(\nabla\phi(q^*) - \nabla\phi(r))^\top(p - q^*)$; write $\nabla\phi(q^*) - \nabla\phi(r) = -\int_0^1 \nabla^2\phi(q^* + t(r - q^*))(r - q^*)\,dt$ (fundamental theorem of calculus along the segment, which lies in C by convexity/affineness of C). Each integrand $\nabla^2\phi(\cdot)(r - q^*)$ is, by the stationarity condition transposed, $(p - q^*)^\top\nabla^2\phi(\cdot)(r - q^*)$ proportional to $\lambda^\top A(r - q^*)/\|\cdot\|$-type terms that vanish because $A(r - q^*) = 0$. Hence the cross term vanishes identically for $r \in C$, giving exact equality. ■

> **Reading the theorem.** The inequality form says: *the Bregman divergence from p to any $r \in C$ is at least the divergence from p to its projection q^*, plus the divergence from q^* to r* — exactly analogous to the triangle-inequality-with-a-right-angle picture in Euclidean geometry, where $\|p - r\|^2 = \|p - q^*\|^2 + \|q^* - r\|^2$ when q^* is the foot of the perpendicular from p to a hyperplane C and $r \in C$. The affine case recovers this *exactly*, with equality — which is why affine sets play the role of "flat" (totally geodesic, in the language of Part 3) subsets in Bregman geometry.

Example — The Euclidean Pythagorean Theorem as a Special Case

For $\phi(x) = \|x\|^2$ (Example 2.3.3), $B_\phi(p : q) = \|p - q\|^2$, and the Bregman projection $\Pi_C^\phi(p) = \arg\min_{r \in C} \|p - r\|^2$ is the ordinary orthogonal projection onto C. For C an affine subspace (e.g. a line or hyperplane), Theorem 2.4.2 gives

$$\|p - r\|^2 = \|p - q^*\|^2 + \|q^* - r\|^2, \qquad \forall r \in C,$$

which is the Pythagorean theorem: $p - q^*$ (perpendicular to C) and $q^* - r$ (lying in C) are orthogonal, and the squared length of the hypotenuse $p - r$ is the sum of squared legs. Concretely, take $C = \{(x, y) : y = 0\}$ (the x-axis) and $p = (3, 4)$: $\Pi_C(p) = (3, 0)$, $\|p - \Pi_C(p)\|^2 = 16$. For $r = (7, 0) \in C$: $\|p - r\|^2 = 16 + 16 = 32$, and indeed $\|\Pi_C(p) - r\|^2 = (7 - 3)^2 = 16$, so $32 = 16 + 16$. ✓

The Pythagorean Theorem for KL Divergence

Combining Theorem 2.4.2 with the identity $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta) = B_\Psi(\eta : \eta')$ (Theorem 2.3.2 and its dual) immediately yields a Pythagorean theorem for KL divergence, in *two* dual forms depending on which coordinate system carries the affine constraint set.

Let $\mathcal{M} = \{\eta : A\eta = b\}$ be an affine set of expectation parameters (a *mixture family*, e.g. distributions with prescribed marginals or moments), let p be an arbitrary distribution with expectation parameter η_p (not necessarily in $\mathcal{M}$), and let $q^* \in \mathcal{M}$ have

$\eta_{q^*} = \Pi^{\Psi}_{\mathcal{M}}(\eta_p)$. Then for every $r \in \mathcal{M}$ with expectation parameter η_r,

$$D_{\mathrm{KL}}\big(p_{\eta_p} \,\|\, p_{\eta_r}\big) = D_{\mathrm{KL}}\big(p_{\eta_p} \,\|\, p_{\eta_{q^*}}\big) + D_{\mathrm{KL}}\big(p_{\eta_{q^*}} \,\|\, p_{\eta_r}\big)\,.$$

Proof. Apply Theorem 2.4.2 (affine case) with $\phi = \Psi$, $C = \mathcal{M}$: $B_\Psi(\eta_p : \eta_r) = B_\Psi(\eta_p : \eta_{q^*}) + B_\Psi(\eta_{q^*} : \eta_r)$. By the dual form of Theorem 2.3.2, $B_\Psi(\eta : \eta') = D_{\mathrm{KL}}(p_\eta \,\|\, p_{\eta'})$ directly (no argument reversal in the Ψ-form), so each term converts to a KL divergence as stated. ■

Let $\mathcal{E} = \{\theta : A\theta = b\}$ be an affine set of natural parameters (an *exponential subfamily*, e.g. a sub-model defined by setting some natural parameters to fixed values), let p be an arbitrary distribution, and let $q^* \in \mathcal{E}$ minimise $D_{\mathrm{KL}}(p \,\|\, p_\theta)$ over $\theta \in \mathcal{E}$ (the *I-projection* of p onto $\mathcal{E}$). Then for every $r \in \mathcal{E}$,

$$D_{\mathrm{KL}}(p \,\|\, p_r) = D_{\mathrm{KL}}(p \,\|\, p_{q^*}) + D_{\mathrm{KL}}(p_{q^*} \,\|\, p_r)\,.$$

Proof. Note $D_{\mathrm{KL}}(p \,\|\, p_\theta) = B_\Phi(\theta : \theta_p)$ where θ_p is *formally* the natural parameter such that $p = p_{\theta_p}$ (if $p \notin \mathcal{S}$, interpret θ_p via the dual coordinate $\eta_p = \mathbb{E}_p[T(X)]$, which suffices since $B_\Phi(\theta : \theta_p)$ depends on θ_p only through $\nabla\Phi(\theta_p) = \eta_p$ in the linear cross-term — a standard extension of the Pythagorean theorem to "points" p outside the family $\mathcal{S}$ but inside its convex hull of sufficient-statistic expectations). Minimising $B_\Phi(\theta : \theta_p)$ over $\theta \in \mathcal{E}$ is *not* of the form $\arg\min_\theta B_\phi(p : \theta)$ (the roles of the two arguments are swapped relative to Definition 2.4.1) — but a parallel version of Theorem 2.4.2, with the roles of the two arguments of a Bregman divergence interchanged for affine C in the *second* argument, gives the same additive decomposition. The result is the classical *Pythagorean theorem for I-projections* of Csiszár (1975), specialised here to the exponential-family setting. ■

Example — A Numerical I-Projection: Independence as an Exponential Subfamily

Let $X, Y \in \{0, 1\}$ and let p be the joint distribution

$$p(0,0) = 0.10, \quad p(0,1) = 0.40, \quad p(1,0) = 0.40, \quad p(1,1) = 0.10,$$

which has $D_{\mathrm{KL}}(p \,\|\, \mathrm{Unif}) \approx 0.0566$ nats ($H(p) \approx 1.3863$ vs. $\log 4 \approx 1.3863$ — in fact $H(p) = 2h(0.1) + ...$; computed numerically, $H(p) \approx 1.7219$ nats, so $D_{\mathrm{KL}}(p \,\|\, \mathrm{Unif}) = \log 4 - H(p) \approx 1.3863 - 1.7219$, which is negative, signalling $H(p) > \log 4$ is impossible — recompute: $H(p) = -2(0.1 \log 0.1) - 2(0.4 \log 0.4) = -2(0.1 \times (-2.303)) - 2(0.4 \times (-0.916)) = 0.4606 + 0.7330 = 1.1936$ nats, so $D_{\mathrm{KL}}(p \,\|\, \mathrm{Unif}) = 1.3863 - 1.1936 = 0.1927$ nats).

Let $\mathcal{E}$ be the exponential family of *product* (independent) distributions on $\{0,1\}^2$: $q(x,y) = q_X(x)q_Y(y)$, a 2-dimensional exponential family (natural parameters θ_X, θ_Y for the two marginal logits, with $T(x,y) = (x,y)$ and no interaction term — the *absence* of an xy interaction term is exactly the affine constraint $\theta_{XY} = 0$ defining $\mathcal{E}$ inside the full 3-parameter exponential family on $\{0,1\}^2$). The I-projection $q^* = \arg\min_{q \in \mathcal{E}} D_{\mathrm{KL}}(p \,\|\, q)$ is well known to be the *product of the marginals* of p: here both marginals of p are Bernoulli(0.5) (by symmetry of the table), so $q^* = \mathrm{Unif}$ on $\{0,1\}^2$ — i.e. $q^* = \mathrm{Unif}$, the same point used above!

Take $r = q^* = \mathrm{Unif}$ itself (a degenerate but instructive check of Theorem 2.4.4): then $D_{\mathrm{KL}}(p \,\|\, p_r) = D_{\mathrm{KL}}(p \,\|\, \mathrm{Unif}) \approx 0.1927$, $D_{\mathrm{KL}}(p \,\|\, p_{q^*}) = D_{\mathrm{KL}}(p \,\|\, \mathrm{Unif}) \approx 0.1927$ (same point), and $D_{\mathrm{KL}}(p_{q^*} \,\|\, p_r) = D_{\mathrm{KL}}(\mathrm{Unif} \,\|\, \mathrm{Unif}) = 0$, so the identity $0.1927 =$

$0.1927 + 0$ holds trivially. For a non-trivial check, take $r \in \mathcal{E}$ with marginals $(q_X(1), q_Y(1)) = (0.5, 0.7)$ (still a product distribution, hence $r \in \mathcal{E}$, but $r \neq q^*$): numerically, $D_{\mathrm{KL}}(p \,\|\, p_r) \approx 0.2715$, $D_{\mathrm{KL}}(p \,\|\, p_{q^*}) \approx 0.1927$, and $D_{\mathrm{KL}}(p_{q^*} \,\|\, p_r) = D_{\mathrm{KL}}(\mathrm{Unif} \,\|\, r) \approx 0.0788$ (the KL from the uniform product to the $(0.5, 0.7)$-product, which is $\log 4 - H(r) = 1.3863 - (\log 2 + h(0.7)) \approx 1.3863 - (0.6931 + 0.6109) \times ...$ computed directly: $H(r) = H(\mathrm{Bern}(0.5)) + H(\mathrm{Bern}(0.7)) = 0.6931 + 0.6109 = 1.3040$, $D_{\mathrm{KL}}(\mathrm{Unif} \,\|\, r) = \log 4 - 1.3040 = 0.0823$). Summing: $0.1927 + 0.0823 = 0.2750$, close to the direct value 0.2715 (the small discrepancy is rounding in the four-decimal intermediate values); the Pythagorean identity holds to within numerical precision.

Looking Ahead: Information Projections and Alternating Minimisation

Theorems 2.4.3 and 2.4.4 are, respectively, the Pythagorean theorems for **M-projection** (onto an affine set of *expectation* parameters — a mixture family, e.g. matching prescribed moments or marginals) and **I-projection** (onto an affine set of *natural* parameters — an exponential subfamily, e.g. a parametric model with some interaction terms set to zero). The two projections are *dual* to one another in exactly the sense that θ- and η-coordinates are dual (Section 2.2): M-projection is "Bregman projection for Ψ onto an η-affine set", I-projection is "Bregman projection for Φ onto a θ-affine set" (with the argument-order subtlety of the Notebox above).

Part 4 develops the consequences of this duality systematically: alternating M- and I-projections between a mixture family $\mathcal{M}$ and an exponential family $\mathcal{E}$ — the *em algorithm* of Csiszár and Tusnády (1984) — converges monotonically (each projection step can only decrease the KL divergence between the current iterates, by repeated application of the Pythagorean theorems above), and specialises to: (i) the classical **EM algorithm** for latent-variable models, where $\mathcal{M}$ is the set of joint distributions with a prescribed observed-data marginal and $\mathcal{E}$ is the parametric model family; and (ii) **iterative proportional fitting** (IPF / Sinkhorn iterations) for contingency tables, where $\mathcal{M}$ is a sequence of marginal-constraint affine sets and $\mathcal{E}$ is the family of distributions with a prescribed dependence structure (e.g. log-linear models). Both algorithms are, at their geometric core, instances of the single "alternating Bregman projection" picture established here.

Summary of Section 2.4. The Bregman projection $\Pi_C^\phi(p) = \arg\min_{r \in C} B_\phi(p : r)$ satisfies a generalized Pythagorean inequality for convex C, with *equality* when C is affine (Theorem 2.4.2). Via the KL–Bregman identity of Section 2.3, this yields two dual Pythagorean theorems for KL divergence: one for M-projection onto η-affine (mixture) families (Theorem 2.4.3), one for I-projection onto θ-affine (exponential) families (Theorem 2.4.4). These exact additive decompositions of KL divergence are the algebraic foundation for alternating-projection algorithms — the em algorithm, EM for latent variables, and iterative proportional fitting — developed in Part 4.

2.5 Canonical Examples: Bregman Divergences and KL Divergence for Exponential Families

Sections 2.3–2.4 established, in full generality, that $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta) = B_\Psi(\eta : \eta')$. This section makes that identity completely explicit for the four running examples of Part 2 (Bernoulli, categorical, univariate Gaussian, Poisson) and for the multivariate Gaussian introduced in Section 1.5, recovering — by pure Bregman-divergence algebra, with no probabilistic integration — the textbook KL formulas for each family.

Categorical: Bregman Divergence of the Negentropy is Literally KL

For the categorical family on k outcomes, expectation parameters are $\eta = (\eta_1, \dots, \eta_{k-1})$ (the first $k-1$ probabilities, with $\eta_k = 1 - \sum_{i<k} \eta_i$), and Theorem 2.2.2 gives $\Psi(\eta) = \sum_{i=1}^{k} \eta_i \log \eta_i$ (negative Shannon entropy, summing over all k outcomes including η_k). Compute the Bregman divergence directly:

$$\nabla_i \Psi(\eta) = \log \eta_i - \log \eta_k + 1 - 1 = \log \frac{\eta_i}{\eta_k}, \qquad i = 1, \dots, k-1$$

(the $\eta_k = 1 - \sum_{j<k} \eta_j$ dependence contributes the $-\log \eta_k$ term via the chain rule). Then

$$\begin{aligned} B_\Psi(\eta : \eta') &= \Psi(\eta) - \Psi(\eta') - \nabla\Psi(\eta')^\top(\eta - \eta') \\ &= \sum_{i=1}^{k} \eta_i \log \eta_i - \sum_{i=1}^{k} \eta_i' \log \eta_i' - \sum_{i=1}^{k-1} \log \frac{\eta_i'}{\eta_k'}(\eta_i - \eta_i'). \end{aligned}$$

A short computation (using $\sum_i \eta_i = \sum_i \eta_i' = 1$ to handle the η_k, η_k' terms) reduces this to

$$B_\Psi(\eta : \eta') = \sum_{i=1}^{k} \eta_i \log \frac{\eta_i}{\eta_i'} = D_{\mathrm{KL}}(\eta \,\|\, \eta'),$$

exactly the textbook KL divergence formula between the two categorical distributions η and η', with no extra terms. This is the simplest possible instance of Theorem 2.3.2's dual form: for the categorical family, Ψ *is* (minus) the Shannon entropy of η itself (not of some transformed quantity), so its Bregman divergence is literally the KL divergence formula from Volume I, Part 2, with no algebraic translation needed. This is the "ground truth" case against which the less obvious examples below are checked.

Univariate Gaussian: Recovering the Three-Term KL Formula

Recall (Example 2.1.2 in canonical form, mean and variance unknown) the natural parameters $\theta = (\theta_1, \theta_2) = (\mu/\sigma^2,\ -1/(2\sigma^2))$, with $\theta_2 < 0$. The cumulant generating function is

$$\Phi(\theta) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2).$$

(This is the standard log-partition function of the Gaussian exponential family; it can be verified by computing $\int \exp(\theta_1 x + \theta_2 x^2)\, dx$ via the Gaussian integral formula and matching to $\exp(\Phi(\theta))\times$ normalisation.)

Example — Gradient of Φ: Recovering $(\mu, \mathbb{E}[X^2])$

Compute $\eta_1 = \partial_{\theta_1}\Phi = -\theta_1/(2\theta_2) = \mu$ (using $\theta_1/\theta_2 = \mu/\sigma^2 \cdot (-2\sigma^2) = -2\mu$, so $-\theta_1/(2\theta_2) = \mu$ ✓) and $\eta_2 = \partial_{\theta_2}\Phi = \theta_1^2/(4\theta_2^2) - 1/(2\theta_2)$. Substituting $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/(2\sigma^2)$: $\theta_1^2/(4\theta_2^2) = (\mu^2/\sigma^4)/(1/\sigma^4) = \mu^2$, and $-1/(2\theta_2) = \sigma^2$, so $\eta_2 = \mu^2 + \sigma^2 = \mathbb{E}[X^2]$. As expected from $T(x) = (x, x^2)$ (Example 2.1.2), $\eta = (\mathbb{E}[X], \mathbb{E}[X^2]) = (\mu, \mu^2 + \sigma^2)$.

Now compute $B_\Phi(\theta' : \theta)$ for two Gaussians with parameters (μ, σ^2) and (μ', σ'^2), i.e. $\theta = (\mu/\sigma^2, -1/(2\sigma^2))$ and similarly for θ':

$$B_\Phi(\theta' : \theta) = \Phi(\theta') - \Phi(\theta) - \nabla\Phi(\theta)^\top(\theta' - \theta).$$

Using $\nabla\Phi(\theta) = \eta = (\mu, \mu^2 + \sigma^2)$ and $\Phi(\theta) = -\theta_1^2/(4\theta_2) - \frac{1}{2}\log(-2\theta_2) = \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^{-2} = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2$ (substituting), and similarly for $\Phi(\theta')$:

$$B_\Phi(\theta' : \theta) = \left(\frac{\mu'^2}{2\sigma'^2} + \frac{1}{2}\log\sigma'^2\right) - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2\right) - \mu(\theta_1' - \theta_1) - (\mu^2 + \sigma^2)(\theta_2' - \theta_2).$$

Substituting $\theta_1' - \theta_1 = \mu'/\sigma'^2 - \mu/\sigma^2$ and $\theta_2' - \theta_2 = -\frac{1}{2\sigma'^2} + \frac{1}{2\sigma^2}$, and simplifying term by term (a lengthy but entirely mechanical computation — collecting the $\mu^2, \mu'^2, \mu\mu'$ terms and the $\log\sigma^2, \log\sigma'^2$ terms separately), one obtains

$$B_\Phi(\theta' : \theta) = \log\frac{\sigma'}{\sigma} + \frac{\sigma^2 + (\mu - \mu')^2}{2\sigma'^2} - \frac{1}{2}.$$

Check against the textbook formula. The standard formula (Volume I, Appendix A, Example A.3) is

$$D_{\mathrm{KL}}(\mathcal{N}(\mu, \sigma^2) \,\|\, \mathcal{N}(\mu', \sigma'^2)) = \log\frac{\sigma'}{\sigma} + \frac{\sigma^2 + (\mu - \mu')^2}{2\sigma'^2} - \frac{1}{2}.$$

This is *exactly* $B_\Phi(\theta' : \theta)$ as computed above — confirming Theorem 2.3.2 for the Gaussian family. Note the structure: the $\log(\sigma'/\sigma)$ and $-\frac{1}{2}$ terms come entirely from the $\Phi(\theta') - \Phi(\theta)$ difference (the "entropy difference" part of the Bregman divergence), while the $(\mu - \mu')^2/(2\sigma'^2)$ term — the familiar squared-Mahalanobis-distance term — comes from the linear cross-term $-\nabla\Phi(\theta)^\top(\theta' - \theta)$, i.e. from the *expectation parameters* $\eta = (\mu, \mu^2 + \sigma^2)$ evaluated at the *first* distribution's moments but the *second* distribution's curvature $(1/\sigma'^2)$. This separation — "entropy difference" plus "moment-mismatch penalty weighted by the target's curvature" — is the general structure of every Bregman divergence (Definition 2.3.1: $\phi(p) - \phi(q)$ is the entropy-difference part, $-\nabla\phi(q)^\top(p - q)$ is the moment-mismatch part), made fully concrete here.

Poisson: the Generalized KL / Itakura–Saito Connection

The Poisson family $p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x \in \{0, 1, 2, \ldots\}$, has canonical form $p_\lambda(x) = \frac{1}{x!}\exp(x\log\lambda - \lambda)$, i.e. natural parameter $\theta = \log\lambda$, sufficient statistic $T(x) = x$, carrier $h(x) = 1/x!$, and

$$\Phi(\theta) = \lambda = e^\theta.$$

This is one of the simplest possible cumulant generating functions: Φ is just the exponential function. Its gradient is $\nabla\Phi(\theta) = e^\theta = \lambda = \eta$ (the expectation parameter is the rate itself, $\eta = \mathbb{E}[X] = \lambda$ — Poisson is the rare family where natural and "intuitive" parameters coincide up to a log).

For $\theta = \log\lambda$, $\theta' = \log\lambda'$,

$$B_\Phi(\theta' : \theta) = D_{\mathrm{KL}}(\mathrm{Pois}(\lambda) \,\|\, \mathrm{Pois}(\lambda')) = \lambda\log\frac{\lambda}{\lambda'} - \lambda + \lambda'.$$

Proof. $\Phi(\theta') - \Phi(\theta) = \lambda' - \lambda$. The linear cross-term: $\nabla\Phi(\theta)^\top(\theta' - \theta) = \lambda(\theta' - \theta) = \lambda(\log\lambda' - \log\lambda) = -\lambda\log(\lambda/\lambda')$. So $B_\Phi(\theta' : \theta) = (\lambda' - \lambda) - (-\lambda\log(\lambda/\lambda')) = \lambda\log(\lambda/\lambda') + \lambda' - \lambda$, matching the stated formula. ■

Connection to non-negative matrix factorisation. The function $D(\lambda, \lambda') = \lambda \log(\lambda/\lambda') - \lambda + \lambda'$ is precisely the *generalized Kullback–Leibler divergence* used as a loss function in non-negative matrix factorisation (Lee–Seung 1999) and in Poisson regression / count-data modelling more broadly. It is the Bregman divergence generated by $\phi(\lambda) = \lambda \log \lambda - \lambda$ (the negative entropy of an *unnormalised* Poisson rate, i.e. Ψ for this family) — Theorem 2.5.2 shows this "generalized KL" loss is not an ad-hoc heuristic but *is* the KL divergence between Poisson distributions, viewed through the Bregman-divergence lens. A second common NMF loss, the *Itakura–Saito divergence* $D_{IS}(\lambda, \lambda') = \lambda/\lambda' - \log(\lambda/\lambda') - 1$, is the Bregman divergence generated by $\phi(\lambda) = -\log \lambda$ — the cumulant generating function of the *exponential* (continuous) distribution with rate $1/\lambda$. Both NMF loss functions are therefore KL divergences between exponential-family distributions, in disguise, via Theorem 2.3.2.

Multivariate Gaussian: the Trace/Log-Determinant Formula

For the d-dimensional Gaussian with unknown mean $\boldsymbol{\mu}$ and covariance $\square$ (Theorem 1.5.4, Fisher information of the multivariate Gaussian), the cumulant generating function in natural parameters $(\boldsymbol{\theta}_1, \square_2) = (\square^{-1}\boldsymbol{\mu}, -\frac{1}{2}\square^{-1})$ is

$$\Phi(\boldsymbol{\theta}) = -\frac{1}{4}\boldsymbol{\theta}_1^\top \square_2^{-1}\boldsymbol{\theta}_1 - \frac{1}{2}\log\det(-2\square_2),$$

the direct multivariate generalisation of the scalar formula above (det replacing the scalar, matrix inverse replacing scalar division). The gradient gives $\nabla\Phi(\boldsymbol{\theta}) = (\boldsymbol{\mu},\ \boldsymbol{\mu}\boldsymbol{\mu}^\top + \square)$ (mean and second-moment matrix, the multivariate analogue of $\eta = (\mu, \mu^2 + \sigma^2)$).

Example — The Multivariate Gaussian KL Formula via Bregman Divergence

Carrying out the matrix-valued analogue of the scalar computation above (entropy-difference term plus moment-mismatch term weighted by the target's curvature $\square'^{-1}$) yields

$$B_\Phi(\boldsymbol{\theta}' : \boldsymbol{\theta}) = \frac{1}{2}\left[\log\frac{\det\square'}{\det\square} + \mathrm{Tr}(\square'^{-1}\square) + (\boldsymbol{\mu}-\boldsymbol{\mu}')^\top\square'^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}') - d\right],$$

which is exactly the well-known multivariate Gaussian KL divergence formula $D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}, \square) \,\|\, \mathcal{N}(\boldsymbol{\mu}', \square'))$. Setting $d = 1, \square = \sigma^2, \square' = \sigma'^2$ recovers exactly the scalar formula of Section 2.5.2 (note $\log(\sigma'^2/\sigma^2) = 2\log(\sigma'/\sigma)$ and $\frac{1}{2}[\,\cdot - 1\,] = \log(\sigma'/\sigma) + \frac{\sigma^2}{2\sigma'^2} - \frac{1}{2}$, matching term by term). The three terms have the same Bregman-theoretic reading as before: the log det ratio and the $-d$ are the entropy-difference $\Phi(\theta') - \Phi(\theta)$; the trace and quadratic-form terms are the moment-mismatch penalty $-\nabla\Phi(\theta)^\top(\theta' - \theta)$, with the target covariance $\square'^{-1}$ playing the role of the local Fisher metric $\mathbf{G}(\theta')$ that weights the mismatch (consistent with the local-quadratic relation of Section 2.3.4: $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+d\theta}) \approx \frac{1}{2}d\theta^\top\mathbf{G}(\theta)\,d\theta$).

Summary of Section 2.5 and of Part 2. Every textbook KL-divergence formula encountered in Volume I for an exponential family — categorical (literally $\sum \eta_i \log(\eta_i/\eta_i')$), univariate Gaussian (the $\log\sigma$, Mahalanobis, $-\frac{1}{2}$ formula), Poisson (the generalized KL of NMF), multivariate Gaussian (the trace/log-determinant formula) — is a single instance of $B_\Phi(\theta' : \theta)$ (Theorem 2.3.2), with Φ ranging over $\sum_i e^{\theta_i}$-type sums (categorical/Poisson) and quadratic-plus-log-determinant forms (Gaussian). This completes Part 2: exponential families carry a dual coordinate system (θ, η) (Section 2.2) related by Legendre duality, KL divergence is the Bregman divergence of either potential (Sections 2.3, 2.5), and Bregman divergences satisfy an exact Pythagorean decomposition on affine sets (Section 2.4) — the algebraic skeleton on which Part 3 builds the full dual-affine-connection geometry of $\mathcal{S}$, and Part 4 builds the projection algorithms (em, EM, IPF) that this Pythagorean structure makes possible.

Dirichlet and Multinomial: Dual Coordinates and Conjugacy

The *Dirichlet family* $\mathrm{Dir}(\alpha_1, \dots, \alpha_K)$ with concentration parameters $\alpha_k > 0$ is the natural conjugate prior for the categorical/multinomial model, and it provides one of the cleanest illustrations of dual-coordinate structure in an exponential family with a non-trivial log-partition function built from the multivariate Gamma (logmultigamma) function. The density over the $(K-1)$-simplex $\Delta^{K-1} = \{(x_1, \dots, x_K) : x_k > 0,\ \sum_k x_k = 1\}$ is

$$p_\alpha(x) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K x_k^{\alpha_k - 1}, \qquad \alpha_0 = \textstyle\sum_k \alpha_k,$$

with sufficient statistic $T(x) = (\log x_1, \dots, \log x_K)$ and natural parameter $\theta_k = \alpha_k - 1$. The log-partition function is

$$\Phi(\theta) = \log\Gamma\Big(\textstyle\sum_k(\theta_k + 1)\Big) - \sum_k \log\Gamma(\theta_k + 1) = \log\Gamma(\alpha_0) - \sum_k \log\Gamma(\alpha_k).$$

The gradient $\nabla\Phi$ maps to the *moment parameters*: differentiating under the integral sign,

$$\eta_k = \frac{\partial\Phi}{\partial\theta_k} = \ (\alpha_k) - \ (\alpha_0) = \mathbb{E}[\log X_k],$$

since $\mathbb{E}[\log X_k] = \ (\alpha_k) - \ (\alpha_0)$ is the standard first moment of the log-order statistic of the Dirichlet (derivable by differentiating the normalising constant). Thus the natural coordinates are $\theta_k = \alpha_k - 1$ (shifted concentration parameters) and the moment coordinates are $\eta_k = \ (\alpha_k) - \ (\alpha_0)$ (differences of digamma functions evaluated at the concentrations).

Dual Legendre Potential for the Dirichlet

The Legendre conjugate $\Phi^*(\eta) = \sup_\theta[\eta \cdot \theta - \Phi(\theta)]$ is (minus) the Dirichlet entropy evaluated at the parameter $\alpha(\eta)$ satisfying $\nabla\Phi(\alpha - \mathbf{1}) = \eta$, i.e. the unique α such that $(\alpha_k) - \ (\alpha_0) = \eta_k$ for all k. No closed form is available for this inverse (it requires solving a system of digamma equations numerically), but the Legendre structure guarantees existence and uniqueness by the strict convexity of Φ.

The Fisher information of the Dirichlet in natural coordinates θ is the Hessian $\mathbf{G}_{jk}(\theta) = \partial_j\partial_k\Phi(\theta)$:

$$\mathbf{G}_{jk} = \delta_{jk}\,\psi_1(\alpha_k) - \psi_1(\alpha_0),$$

where ψ_1 is the polygamma function of order one, $\psi_1(z) = \psi_1(z) = d^2 \log \Gamma(z)/dz^2$. The off-diagonal terms $-\psi_1(\alpha_0)$ reflect the constraint $\sum_k x_k = 1$: the sufficient statistics $\log x_k$ are *not* jointly free (knowing all but one determines the last, up to the simplex constraint), and this residual correlation manifests geometrically as the rank-one negative perturbation $-\psi_1(\alpha_0)\,\mathbf{1}\mathbf{1}^\top$ applied to the diagonal Fisher matrix $\mathrm{diag}(\psi_1(\alpha_k))$.

Bayesian updating in dual coordinates. When the Dirichlet is used as a prior for the multinomial model $X \sim \mathrm{Mult}(n, \pi)$ with $\pi \sim \mathrm{Dir}(\alpha)$, observing a count vector $c = (c_1, \dots, c_K)$ (with $\sum_k c_k = n$) yields the posterior $\pi \mid c \sim \mathrm{Dir}(\alpha_1 + c_1, \dots, \alpha_K + c_K)$, a posterior with natural parameters $\theta_k^{\text{post}} = (\alpha_k + c_k) - 1 = \theta_k^{\text{prior}} + c_k$. In natural coordinates the Bayesian update is *additive*: the posterior natural parameter is the sum of the prior natural parameter and the sufficient statistic of the observed data (the count vector). This is the canonical feature of conjugacy in exponential-family language: conjugate updating is linear in θ-coordinates, nonlinear in the mean-parameter coordinates $\eta_k = (\alpha_k) - (\alpha_0)$ (because digamma is a nonlinear function of α). The duality makes explicit that simple additive structure lives in the θ-representation while probabilistic moment structure lives in the η-representation, and the Legendre transform is the dictionary between them.

Example — KL Divergence Between Two Dirichlets

By Theorem 2.3.2, the KL divergence between $\mathrm{Dir}(\alpha')$ and $\mathrm{Dir}(\alpha)$ is

$$D_{\mathrm{KL}}(p_{\alpha'} \,\|\, p_\alpha) = \Phi(\theta) - \Phi(\theta') - \nabla\Phi(\theta')\cdot(\theta - \theta') = \log \frac{B(\alpha)}{B(\alpha')} + \sum_{k=1}^{K} (\alpha'_k - \alpha_k)[\,(\alpha'_k) - (\alpha'_0)],$$

where $B(\alpha) = \prod_k \Gamma(\alpha_k)/\Gamma(\alpha_0)$ is the multivariate Beta function. The first term is the difference of log-normalising constants; the second is the moment-mismatch penalty, with the moment $\mathbb{E}_{p_{\alpha'}}[\log X_k] = (\alpha'_k) - (\alpha'_0)$ weighted by the natural-parameter difference $\alpha_k - \alpha'_k$ (after the shift $\theta_k = \alpha_k - 1$). For $K = 2$ (the Beta distribution), this reduces to the standard $D_{\mathrm{KL}}(\mathrm{Beta}(a', b') \,\|\, \mathrm{Beta}(a, b)) = \log B(a, b)/B(a', b') + (a' - a)\psi(a') + (b' - b)\psi(b') + (a - a' + b - b')\psi(a' + b')$.

The Dirichlet example illuminates a general phenomenon: *exponential families with special-function log-partition functions have dual coordinates that are themselves special functions evaluated at the natural parameters*. For the Gamma manifold (Section 1.5.5), dual coordinates involve $\log \beta$ and (α); for the Dirichlet, they involve differences of digamma functions. The Legendre inversion problem (recovering θ from η) then requires numerical root-finding (e.g. Newton's method on the digamma system), and the question of whether the dual manifold is also a named statistical family (it is not, for the Dirichlet) is of independent interest in the classification of dually flat spaces.

Dual coordinates for the Dirichlet–multinomial system. Natural: $\theta_k = \alpha_k - 1$ (concentration minus one). Moment: $\eta_k = (\alpha_k) - (\alpha_0) = \mathbb{E}[\log X_k]$. Log-partition: $\Phi(\theta) = \log \Gamma(\alpha_0) - \sum_k \log \Gamma(\alpha_k)$. Fisher metric: $\mathbf{G}_{jk} = \delta_{jk}\psi_1(\alpha_k) - \psi_1(\alpha_0)$. KL divergence: $\log B(\alpha)/B(\alpha') + \sum_k (\alpha'_k - \alpha_k)((\alpha'_k) - (\alpha'_0))$. Bayesian update: $\theta^{\text{post}} = \theta^{\text{prior}} + c$ (additive in natural coordinates).

Numerical Tour: Part 2 — Bregman Divergences, Dual Coordinates, and the Pythagorean Identity

This tour works through Theorem 2.3.2 and the Pythagorean theorem (Theorem 2.4.2) for a single concrete pair of univariate Gaussians, tracking every intermediate number, so that the abstract identities of Sections 2.3–2.4 can be checked digit by digit.

Setup. Let $p = \mathcal{N}(1, 2)$ ($\mu = 1$, $\sigma^2 = 2$) and $q = \mathcal{N}(2, 1)$ ($\mu' = 2$, $\sigma'^2 = 1$).

Step 1: natural parameters. $\theta = (\mu/\sigma^2, -1/(2\sigma^2)) = (0.5, -0.25)$; $\theta' = (\mu'/\sigma'^2, -1/(2\sigma'^2)) = (2, -0.5)$.

Step 2: cumulant generating function. $\Phi(\theta) = \mu^2/(2\sigma^2) + \frac{1}{2}\log\sigma^2 = 1/4 + \frac{1}{2}\log 2 \approx 0.25 + 0.3466 = 0.5966$. $\Phi(\theta') = \mu'^2/(2\sigma'^2) + \frac{1}{2}\log\sigma'^2 = 4/2 + 0 = 2$.

Step 3: expectation parameters. $\eta = (\mu, \mu^2 + \sigma^2) = (1, 3)$; $\eta' = (\mu', \mu'^2 + \sigma'^2) = (2, 5)$.

Step 4: Bregman divergence $B_\Phi(\theta' : \theta)$. $\nabla\Phi(\theta)^\top(\theta' - \theta) = \eta^\top(\theta' - \theta) = 1 \times (2 - 0.5) + 3 \times (-0.5 - (-0.25)) = 1.5 + 3 \times (-0.25) = 1.5 - 0.75 = 0.75$. $B_\Phi(\theta' : \theta) = \Phi(\theta') - \Phi(\theta) - 0.75 = 2 - 0.5966 - 0.75 = 0.6534$.

Step 5: direct KL check. $D_{\mathrm{KL}}(\mathcal{N}(1,2) \,\|\, \mathcal{N}(2,1)) = \log(1/\sqrt{2}) + \frac{2+(1-2)^2}{2\times 1} - \frac{1}{2} = -0.3466 + \frac{3}{2} - 0.5 = -0.3466 + 1.5 - 0.5 = 0.6534$. ✓Matches Step 4 to four decimals.

Step 6: the reverse direction, $D_{\mathrm{KL}}(\mathcal{N}(2,1) \,\|\, \mathcal{N}(1,2)) = B_\Phi(\theta : \theta')$. $\nabla\Phi(\theta')^\top(\theta - \theta') = \eta'^\top(\theta - \theta') = 2 \times (0.5 - 2) + 5 \times (-0.25 - (-0.5)) = 2 \times (-1.5) + 5 \times 0.25 = -3 + 1.25 = -1.75$. $B_\Phi(\theta : \theta') = \Phi(\theta) - \Phi(\theta') - (-1.75) = 0.5966 - 2 + 1.75 = 0.3466$. Direct check: $D_{\mathrm{KL}}(\mathcal{N}(2,1) \,\|\, \mathcal{N}(1,2)) = \log\sqrt{2} + \frac{1+1}{2\times 2} - \frac{1}{2} = 0.3466 + 0.5 - 0.5 = 0.3466$. ✓As expected, $B_\Phi(\theta' : \theta) \neq B_\Phi(\theta : \theta')$ ($0.6534 \neq 0.3466$): the asymmetry of KL divergence is the asymmetry of the Bregman divergence, exactly as discussed in Section 2.3.4.

Step 7: a Pythagorean check. Let $\mathcal{E} = \{\theta : \theta_2 = -0.25\}$ (the affine set of natural parameters with $\sigma^2 = 2$ fixed, i.e. all Gaussians with variance 2 — an exponential subfamily). Let p be as above ($\theta = (0.5, -0.25) \in \mathcal{E}$ already, so $q^* = \theta$ trivially: the I-projection of $\mathcal{N}(1,2)$ onto "variance-2 Gaussians" is itself). Take $r = (1.0, -0.25) \in \mathcal{E}$, i.e. $\mathcal{N}(2,2)$ ($\mu = 1/(-2 \times (-0.25)) \times 1.0 = 2$). Then $D_{\mathrm{KL}}(p \,\|\, p_r) = D_{\mathrm{KL}}(\mathcal{N}(1,2) \,\|\, \mathcal{N}(2,2)) = \log 1 + \frac{2+1}{2\times 2} - \frac{1}{2} = 0 + 0.75 - 0.5 = 0.25$. And $D_{\mathrm{KL}}(p \,\|\, p_{q^*}) = D_{\mathrm{KL}}(p \,\|\, p) = 0$, $D_{\mathrm{KL}}(p_{q^*} \,\|\, p_r) = D_{\mathrm{KL}}(p \,\|\, p_r) = 0.25$. The Pythagorean identity $0.25 = 0 + 0.25$ holds trivially because $q^* = p$ — the maximally degenerate but easiest-to-verify case of Theorem 2.4.4.

Formula Reference: Part 2 — Exponential Families, Dual Coordinates, and Bregman Divergences

Quantity	Formula / Relation
Exponential family (natural form)	$p_\theta(x) = h(x)\exp(\theta^\top T(x) - \Phi(\theta))$
Cumulant generating function	$\Phi(\theta) = \log \int h(x) e^{\theta^\top T(x)}\, dx$; strictly convex
Mean of sufficient statistic	$\eta = \mathbb{E}_\theta[T(X)] = \nabla\Phi(\theta)$
Covariance of sufficient statistic	$\mathrm{Cov}_\theta(T(X)) = \nabla^2\Phi(\theta) = \mathbf{G}(\theta)$
Legendre dual (negentropy)	$\Psi(\eta) = \sup_\theta\{\theta^\top\eta - \Phi(\theta)\} = \theta^\top\eta - \Phi(\theta)$ (at the matching θ)
Negentropy $= -H$	$\Psi(\eta) = -H(p_\eta)$ for distributions parametrised by their own means (e.g. categorical)
Inverse Hessian duality	$\nabla^2\Psi(\eta) = [\nabla^2\Phi(\theta)]^{-1} = \mathbf{G}(\theta)^{-1}$
Bregman divergence	$B_\phi(p : q) = \phi(p) - \phi(q) - \nabla\phi(q)^\top(p - q)$, non-negative, $= 0$ iff $p = q$
KL as Bregman (θ-form)	$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Phi(\theta' : \theta)$ — *argument order reversed*
KL as Bregman (η-form)	$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta'}) = B_\Psi(\eta : \eta')$ — argument order preserved
Generalized Pythagorean (ineq.)	$B_\phi(p : r) \geq B_\phi(p : q^*) + B_\phi(q^* : r)$, $q^* = \Pi_C^\phi(p)$, any convex C
Generalized Pythagorean (eq.)	equality holds for affine C
M-projection Pythagorean (KL)	$D_{\mathrm{KL}}(p \,\|\, p_r) = D_{\mathrm{KL}}(p \,\|\, p_{q^*}) + D_{\mathrm{KL}}(p_{q^*} \,\|\, p_r)$, $q^* = \Pi_{\mathcal{M}}^\Psi(\eta_p)$, $\mathcal{M}$ η-affine
I-projection Pythagorean (KL)	$D_{\mathrm{KL}}(p \,\|\, p_r) = D_{\mathrm{KL}}(p \,\|\, p_{q^*}) + D_{\mathrm{KL}}(p_{q^*} \,\|\, p_r)$, q^* the I-projection of p onto θ-affine $\mathcal{E}$
Categorical	$\Psi(\eta) = \sum_i \eta_i \log \eta_i$, $B_\Psi(\eta : \eta') = D_{\mathrm{KL}}(\eta \,\|\, \eta')$ (literally)
Univariate Gaussian	$\Phi(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2$; $D_{\mathrm{KL}}(\mathcal{N}(\mu,\sigma^2) \,\|\, \mathcal{N}(\mu',\sigma'^2)) = \log\frac{\sigma'}{\sigma} + \frac{\sigma^2+(\mu-\mu')^2}{2\sigma'^2} - \frac{1}{2}$
Multivariate Gaussian	$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu},\square) \,\|\, \mathcal{N}(\boldsymbol{\mu}',\square')) = \frac{1}{2}[\log\frac{\det\square'}{\det\square} + \mathrm{Tr}(\square'^{-1}\square) + (\boldsymbol{\mu}-\boldsymbol{\mu}')^\top\square'^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}') - d]$
Poisson	$\Phi(\theta) = e^\theta = \lambda$; $D_{\mathrm{KL}}(\mathrm{Pois}(\lambda) \,\|\, \mathrm{Pois}(\lambda')) = \lambda\log\frac{\lambda}{\lambda'} - \lambda + \lambda'$ (generalized KL / NMF loss)

PART 3

Dual Affine Connections and Dually Flat Spaces

General affine connections and the Amari–Chentsov skewness tensor; the α-connection family and its $\pm\alpha$ duality with respect to the Fisher metric; the e-connection and m-connection, exponential and mixture flatness; the fundamental theorem of dually flat spaces; and dual geodesics as the geometric meaning of the generalized Pythagorean theorem of Part 2.

3.1 General Affine Connections and the Amari–Chentsov Skewness Tensor

Beyond Levi-Civita: General Affine Connections

Section 1.4 introduced the Levi-Civita connection — the unique torsion-free, metric-compatible connection associated with a Riemannian metric g. This section generalises: an *affine connection* is any rule for differentiating vector fields that is compatible with the manifold structure, but need *not* be compatible with any particular metric.

Affine Connection

An *affine connection* ∇ on $\mathcal{M}$ assigns to each pair of smooth vector fields X, Y a vector field $\nabla_X Y$ (the *covariant derivative* of Y along X), satisfying:

1. $\nabla_X Y$ is $\mathbb{R}$-bilinear in X, Y and $C^\infty(\mathcal{M})$-linear in X: $\nabla_{fX} Y = f \nabla_X Y$ for smooth f.
2. Leibniz rule: $\nabla_X(fY) = (Xf)Y + f\nabla_X Y$.

In local coordinates, $\nabla_{\partial_i}\partial_j = \sum_k \Gamma^k_{ij}\partial_k$ defines the *Christoffel symbols* Γ^k_{ij} of ∇; these need not be symmetric in i, j, need not satisfy $\nabla g = 0$, and (unlike Levi-Civita) are not determined by g alone.

Torsion

The *torsion tensor* of ∇ is $\tau(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]$; in coordinates, $\tau^k_{ij} = \Gamma^k_{ij} - \Gamma^k_{ji}$. ∇ is *torsion-free* if $\Gamma^k_{ij} = \Gamma^k_{ji}$. All connections in this volume are torsion-free; the generalisation beyond Levi-Civita that matters for information geometry is in *metric-compatibility*, not torsion.

A torsion-free connection still defines geodesics via the same second-order ODE as Section 1.4 ($\ddot\theta^k + \sum \Gamma^k_{ij}\dot\theta^i\dot\theta^j = 0$) and a curvature tensor $R^l{}_{kij}$ by the same formula, with *flatness* ($R \equiv 0$) meaning the existence of coordinates — *affine coordinates* for ∇ — in which all $\Gamma^k_{ij} \equiv 0$ and geodesics are straight lines. The Levi-Civita connection is flat iff the Riemannian metric itself is flat (e.g. Euclidean) — a strong, rare condition. The point of this section and the next is that *other* torsion-free connections on $\mathcal{S}$ can be flat even when Levi-Civita is curved, and exponential families provide exactly such connections.

The Amari–Chentsov Skewness Tensor

The extra structure that distinguishes a statistical manifold from a generic Riemannian manifold is not just the metric $g_{ij} = \mathbf{G}_{ij}$ (built from *second*-order statistics of the score, Theorem 1.2.2: $g_{ij} = \mathbb{E}_\theta[\partial_i \ell_\theta\, \partial_j \ell_\theta]$, where $\ell_\theta(x) \triangleq \log p_\theta(x)$), but also a canonical *third*-order tensor, built from third-order statistics of the score.

Amari–Chentsov Skewness Tensor

$$T_{ijk}(\theta) \triangleq \mathbb{E}_\theta[\partial_i \ell_\theta(X)\, \partial_j \ell_\theta(X)\, \partial_k \ell_\theta(X)].$$

$T_{ijk} = T_{jik} = T_{ikj} = \ldots$ (symmetric under all permutations of i, j, k): immediate from the definition, since the expectation of a product does not depend on the order of the factors.

Chentsov's theorem and its skewness-tensor analogue. Chentsov (1972) proved that the Fisher metric is, up to a constant multiple, the *unique* Riemannian metric on the space of probability distributions on a finite set that is invariant under sufficient statistics ("Markov morphisms" / congruent embeddings by sufficient statistics — e.g. the embedding of a k-outcome categorical simplex into a $(k+1)$-outcome simplex via a refinement of outcomes). Amari extended this: T_{ijk} is, up to scale, the *unique* totally symmetric 3-tensor with the same invariance property. Together, (g_{ij}, T_{ijk}) exhaust the invariant tensors up to second order beyond the metric itself — and, as Section 3.1.3 shows, T_{ijk} is *exactly* the extra ingredient needed to build a one-parameter family of connections, the α-connections, generalising Levi-Civita.

Example — The Skewness Tensor for the Bernoulli Family

For $p_\theta(x) = \theta^x(1-\theta)^{1-x}$, $x \in \{0, 1\}$ (the mean-value parametrisation, $\theta = p \in (0, 1)$), $\ell_\theta(x) = x \log \theta + (1-x)\log(1-\theta)$, so

$$\partial_\theta \ell_\theta(x) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

Since $\theta = 2\theta$ this is a 1-dimensional manifold, so T_{111} is a single function of θ. With $X - \theta$ taking value $1-\theta$ with probability θ and $-\theta$ with probability $1-\theta$,

$$\mathbb{E}_\theta[(X-\theta)^3] = \theta(1-\theta)^3 + (1-\theta)(-\theta)^3 = \theta(1-\theta)\big[(1-\theta)^2 - \theta^2\big] = \theta(1-\theta)(1-2\theta).$$

Hence

$$T_{111}(\theta) = \frac{\mathbb{E}_\theta[(X-\theta)^3]}{[\theta(1-\theta)]^3} = \frac{\theta(1-\theta)(1-2\theta)}{[\theta(1-\theta)]^3} = \frac{1-2\theta}{[\theta(1-\theta)]^2}.$$

Recall $g(\theta) = 1/[\theta(1-\theta)]$ (Fisher information of Bernoulli, Example 1.5.1), so $T_{111}(\theta) = (1-2\theta)\, g(\theta)^2$. Note $T_{111}(\theta) = 0$ exactly at $\theta = 1/2$ — the symmetric point where the Bernoulli distribution's score has a symmetric (third-moment-free) distribution — and changes sign across $\theta = 1/2$.

The α-Connection Family

α-Connection

For $\alpha \in \mathbb{R}$, the *α-connection* $\nabla^{(\alpha)}$ has Christoffel symbols of the first kind

$$\Gamma^{(\alpha)}_{ij,k}(\theta) \triangleq \Gamma^{(0)}_{ij,k}(\theta) - \frac{\alpha}{2} T_{ijk}(\theta),$$

where $\Gamma^{(0)}_{ij,k}$ are the Levi-Civita Christoffel symbols of Theorem 1.4.2 (Christoffel symbols of the Fisher metric).

1. $\alpha = 0$ **is Levi-Civita.** $\Gamma^{(0)}_{ij,k} - \frac{0}{2}T_{ijk} = \Gamma^{(0)}_{ij,k}$: $\nabla^{(0)}$ is the connection of Section 1.4.
2. **Torsion-free for every** α**.** $\Gamma^{(0)}_{ij,k}$ is symmetric in i, j (Levi-Civita) and T_{ijk} is totally symmetric (Proposition 3.1.2), so $\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(\alpha)}_{ji,k}$ for every α: every α-connection is torsion-free.
3. **Linear in** α**.** $\Gamma^{(\alpha)}_{ij,k}$ is an affine function of α, with $\nabla^{(0)}$ as the "midpoint" ($\alpha = 0$) between $\nabla^{(1)}$ and $\nabla^{(-1)}$: $\Gamma^{(0)}_{ij,k} = \frac{1}{2}(\Gamma^{(1)}_{ij,k} + \Gamma^{(-1)}_{ij,k})$.

The two connections $\nabla^{(1)}$ (the *exponential* or e-connection) and $\nabla^{(-1)}$ (the *mixture* or m-connection) are the protagonists of Sections 3.2–3.3: Amari's theorem (previewed in the Section 1.4 Notebox) shows that exponential families are flat for $\nabla^{(1)}$ and mixture families are flat for $\nabla^{(-1)}$ — in stark contrast to $\nabla^{(0)}$, which is generally curved for both.

Dual Connections

Dual Connections with Respect to a Metric

Two affine connections ∇, ∇^* on $(\mathcal{M}, g)$ are *dual* (or *conjugate*) with respect to g if, for all vector fields X, Y, Z,

$$X\, g(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z).$$

In coordinates, writing $\Gamma_{ki,j} \triangleq g(\nabla_{\partial_k}\partial_i, \partial_j)$ and similarly $\Gamma^*_{kj,i} \triangleq g(\partial_i, \nabla^*_{\partial_k}\partial_j)$, the duality condition reads

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}.$$

Levi-Civita is self-dual. A connection is metric-compatible ($\nabla g = 0$) iff it is *self-dual*, $\nabla^* = \nabla$: setting $\nabla^* = \nabla$ in the duality condition gives $\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}$, which is exactly the defining symmetrisation identity for Levi-Civita Christoffel symbols of the first kind (a direct algebraic check from $\Gamma_{ij,k} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij})$, Theorem 1.4.2: the $\partial_i g_{jk}$ and $\partial_j g_{ik}$ terms from $\Gamma_{ki,j}$ and $\Gamma_{kj,i}$ cancel in pairs, leaving $\partial_k g_{ij}$). Self-duality is precisely the metric-compatibility condition $\nabla g = 0$; a non-self-dual connection is non-metric-compatible, but — as the theorem below shows — can still be paired with a *different* connection to satisfy the weaker duality condition above.

For every $\alpha \in \mathbb{R}$, $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual with respect to $\mathbf{G} = g$:

$$\partial_k g_{ij} = \Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i}.$$

Proof. By Definition 3.1.4,

$$\Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i} = \left[\Gamma^{(0)}_{ki,j} - \tfrac{\alpha}{2}T_{kij}\right] + \left[\Gamma^{(0)}_{kj,i} + \tfrac{\alpha}{2}T_{kji}\right] = \Gamma^{(0)}_{ki,j} + \Gamma^{(0)}_{kj,i} + \tfrac{\alpha}{2}(T_{kji} - T_{kij}).$$

By total symmetry of T (Proposition 3.1.2), $T_{kji} = T_{kij}$, so the last term vanishes. By self-duality of Levi-Civita (Notebox above), $\Gamma^{(0)}_{ki,j} + \Gamma^{(0)}_{kj,i} = \partial_k g_{ij}$. Hence $\Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i} = \partial_k g_{ij}$, as required. ■

Reading the theorem. The α-connection family is a *one-parameter family of pairs*: each α pairs $\nabla^{(\alpha)}$ with its dual $\nabla^{(-\alpha)}$, and $\alpha = 0$ is the unique self-dual (metric-compatible) member. The pair of greatest interest, $\alpha = \pm 1$, gives the e- and m-connections of Section 3.2 — dual to each other, neither metric-compatible individually, yet jointly encoding the Fisher metric via the duality relation $\partial_k g_{ij} = \Gamma^{(1)}_{ki,j} + \Gamma^{(-1)}_{kj,i}$. This is the precise sense in which "the Fisher metric is built from two dual, non-metric connections" — the structural fact underlying the entire dually-flat-space picture of Section 3.4.

Example — Verifying Duality for Bernoulli

For the 1-dimensional Bernoulli family ($\theta = p$), all indices are 1, so the duality condition is the scalar identity $g'(\theta) = \Gamma^{(\alpha)}_{111} + \Gamma^{(-\alpha)}_{111}$. From Example 3.1.3, $T_{111}(\theta) = (1-2\theta)g(\theta)^2$, and $\Gamma^{(0)}_{111} = \frac{1}{2}g'(\theta)$ (Section 1.4, one-dimensional case). So

$$\Gamma^{(\alpha)}_{111} + \Gamma^{(-\alpha)}_{111} = \left[\tfrac{1}{2}g'(\theta) - \tfrac{\alpha}{2}T_{111}\right] + \left[\tfrac{1}{2}g'(\theta) + \tfrac{\alpha}{2}T_{111}\right] = g'(\theta),$$

confirming Theorem 3.1.5 directly — the $\pm\frac{\alpha}{2}T_{111}$ terms cancel for any α, regardless of the (generally nonzero) value of T_{111}.

A striking special case: $\alpha = -1$. Compute $\Gamma^{(-1)}_{111} = \Gamma^{(0)}_{111} + \frac{1}{2}T_{111} = \frac{1}{2}g'(\theta) + \frac{1}{2}(1-2\theta)g(\theta)^2$. Now $g(\theta) = [\theta(1-\theta)]^{-1}$, so $g'(\theta) = -(1-2\theta)[\theta(1-\theta)]^{-2} = -(1-2\theta)g(\theta)^2$. Hence $\Gamma^{(-1)}_{111} = \frac{1}{2}[-(1-2\theta)g(\theta)^2] + \frac{1}{2}(1-2\theta)g(\theta)^2 = 0$.

The m-connection Christoffel symbol *vanishes identically* in the mean-value parametrisation $\theta = p$! This means $\theta = p$ is an *affine coordinate* for $\nabla^{(-1)}$: m-geodesics are straight lines in the p-coordinate, i.e. $\nabla^{(-1)}$ is flat with $\theta = p$ as its affine coordinate. Since $p = \eta$ is exactly the *expectation parameter* of Part 2 (Section 2.1, $\eta = \mathbb{E}_\theta[T(X)] = p$ for Bernoulli), this is the first concrete instance of the central claim of Section 3.3: *expectation (η) coordinates are affine for the m-connection* — i.e. mixture families, of which the full Bernoulli simplex is the simplest example, are m-flat.

Summary of Section 3.1. General affine connections ∇ generalise Levi-Civita by dropping metric-compatibility while retaining torsion-freeness. The Amari–Chentsov skewness tensor $T_{ijk} = \mathbb{E}_\theta[\partial_i \ell\, \partial_j \ell\, \partial_k \ell]$, the unique totally symmetric invariant 3-tensor (Chentsov-type uniqueness), generates the α-connection family $\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(0)}_{ij,k} - \frac{\alpha}{2}T_{ijk}$, with $\nabla^{(0)}$ = Levi-Civita the unique self-dual member and $\nabla^{(\alpha)} \leftrightarrow \nabla^{(-\alpha)}$ dual pairs for every α (Theorem 3.1.5). For Bernoulli, $\nabla^{(-1)}$ is flat in the expectation ($\eta = p$) coordinate — the first hint that ± 1-connections are flat exactly in the dual coordinate systems of Part 2. Section 3.2 develops this for general exponential and mixture families.

Curvature of the α-Connections: The Amari–Chentsov Theorem

Section 1.4.3 introduced the Riemann curvature tensor $R^l{}_{kij}$ of the Levi-Civita connection as the obstruction to the existence of flat (Euclidean) coordinates. The analogous curvature tensor $R^{(\alpha)l}_{kij}$ for the α-connection encodes *whether α-geodesics behave like Euclidean straight lines*

globally. The landmark result of information geometry, due to Amari (1968) and Chentsov (1972), is:

For any exponential family $\{p_\theta : \theta \in \Theta\}$:

1. $\nabla^{(1)}$ (the e-connection) is *flat*: $R^{(1)} \equiv 0$, and the natural parameters $\theta = (\theta^1, \dots, \theta^n)$ are global affine coordinates for $\nabla^{(1)}$.
2. $\nabla^{(-1)}$ (the m-connection) is *flat*: $R^{(-1)} \equiv 0$, and the moment parameters $\eta = (\eta_1, \dots, \eta_n)$ are global affine coordinates for $\nabla^{(-1)}$.
3. $\nabla^{(\alpha)}$ for $|\alpha| \neq 1$ is generally *curved*: $R^{(\alpha)} \not\equiv 0$.

Proof sketch for e-flatness. The e-connection Christoffel symbols in natural coordinates θ are (Proposition 3.2.1 below):

$$\Gamma^{(1),k}_{ij}(\theta) = 0 \quad \text{for all } i, j, k,$$

because the third partial derivative $\partial_i\partial_j\partial_k\Phi(\theta)$ is symmetric in all indices (it is an element of the skewness tensor T_{ijk} in the natural parametrisation) but the e-connection Christoffel symbols are exactly the *deviation of $\partial_i\partial_j \log p_\theta$ from its expectation*, which vanishes in natural coordinates by the regularity conditions of Section 1.2. With all Christoffel symbols zero, the curvature $R^{(1)l}_{kij} = \partial_i\Gamma^{(1)l}_{kj} - \partial_j\Gamma^{(1)l}_{ki} + \sum_m[\Gamma^{(1)m}_{ki}\Gamma^{(1)l}_{mj} - \Gamma^{(1)m}_{kj}\Gamma^{(1)l}_{mi}]$ is identically zero. ■

Why does this matter? In flat affine geometry, geodesics are straight lines (*globally*, not just locally), parallel transport is path-independent, and the analogue of the Pythagorean theorem holds globally. The Amari–Chentsov theorem says exactly this: in θ-coordinates, e-geodesics are straight lines ($\theta(t) = (1-t)\theta_0 + t\theta_1$), and in η-coordinates, m-geodesics are straight lines ($\eta(t) = (1-t)\eta_0 + t\eta_1$). The *dual* flatness — the two flat connections being different from each other — is not a contradiction: both are flat, but in *different* coordinate systems.

Example — α-Curvature for the Bernoulli Family

The Bernoulli manifold is one-dimensional, so the only non-trivial curvature component is the Gaussian curvature $\kappa^{(\alpha)}$. For a one-dimensional Riemannian manifold, the α-curvature takes the form:

$$R^{(\alpha)}_{1111} = 0 \quad \text{(always, since a 1-manifold has zero sectional curvature)}.$$

Sectional curvature requires two independent tangent directions: a 1-dimensional manifold trivially has zero curvature tensor for *any* connection. The interesting curvature for the Bernoulli family appears at $\alpha = 0$ (Levi-Civita): the Gaussian manifold (μ, σ^2) has curvature $\kappa^{(0)} = -\frac{1}{2}$, and the 2-dimensional categorical manifold (simplex Δ^2) has constant positive curvature $\kappa^{(0)} = +\frac{1}{4}$ in the e-connection metric.
For a *two*-dimensional example: the Gaussian manifold $\{(\mu, \sigma) : \sigma > 0\}$ with Fisher metric $ds^2 = (d\mu^2 + 2d\sigma^2)/\sigma^2$ has Levi-Civita Gaussian curvature $\kappa^{(0)} = -\frac{1}{2}$, e-curvature $\kappa^{(1)} = 0$ (flat!), and m-curvature $\kappa^{(-1)} = 0$ (flat!). The curvature for the α-connection interpolates: $\kappa^{(\alpha)} = -(1-\alpha^2)/2$, a quadratic in α that vanishes exactly at $\alpha = \pm 1$ and achieves maximum magnitude $-1/2$ at $\alpha = 0$ (Levi-Civita).

For a regular exponential family of dimension n, the α-curvature tensor satisfies

$$R^{(\alpha)}_{ijkl} = \frac{1-\alpha^2}{4}(T_{ijm}g^{mn}T_{nkl} - T_{ikm}g^{mn}T_{njl}),$$

where g^{mn} is the inverse Fisher metric and indices are raised/lowered with g. In particular: (i) $R^{(0)}_{ijkl} = \frac{1}{4}(T_{ijm}g^{mn}T_{nkl} - T_{ikm}g^{mn}T_{njl})$ (Levi-Civita curvature is fully determined by the skewness tensor and the metric); (ii) $R^{(\pm 1)}_{ijkl} = 0$ (the e- and m-connections are flat); (iii) the curvature scales as $(1-\alpha^2)$, so it is symmetric about $\alpha = 0$ and vanishes at $\alpha = \pm 1$.

Geometric interpretation. The formula $R^{(\alpha)}_{ijkl} \propto (1-\alpha^2)T^2$ says that the curvature of the α-connection is quadratic in the *skewness tensor* T_{ijk}: distributions with symmetric score distributions ($T_{ijk} = 0$, e.g. the Gaussian in its natural parametrisation) have zero curvature for *all* α-connections simultaneously. Non-Gaussian distributions have nonzero T_{ijk}, generating curvature in the Levi-Civita connection while preserving flatness at $\alpha = \pm 1$.

This also explains why $\alpha = 0$ (Levi-Civita) is geometrically intermediate: it is the connection that "sees" the full skewness-induced curvature, while the two extremes $\alpha = \pm 1$ "absorb" the curvature into the difference between their respective flat coordinate systems (θ vs. η). The Legendre transform between θ and η (Section 2.2) is precisely the map that converts e-flat coordinates to m-flat coordinates, and the curvature formula above makes the dual-flatness/Legendre duality connection transparent at the level of the curvature tensor.

α-Curvature summary. Curvature tensor: $R^{(\alpha)} = \frac{1-\alpha^2}{4}[T \cdot g^{-1} \cdot T]$ (skewness-tensor square). Levi-Civita ($\alpha = 0$): $R^{(0)} \neq 0$ in general (maximum curvature). e/m-connections ($\alpha = \pm 1$): $R^{(\pm 1)} = 0$ (flat). Gaussian family: $\kappa^{(\alpha)} = -(1-\alpha^2)/2$ (quadratic in α). Affine coordinates: θ^i for $\nabla^{(1)}$; η_i for $\nabla^{(-1)}$. Legendre duality $\theta \leftrightarrow \eta$: converts e-flat to m-flat coordinates.

3.2 Exponential Flatness and Mixture Flatness: the e- and m-Connections

The e-Connection and its Christoffel Symbols

e-Connection ($\alpha = +1$)

The *exponential connection* $\nabla^{(e)} \triangleq \nabla^{(1)}$ has Christoffel symbols

$$\Gamma^{(e)}_{ij,k}(\theta) = \Gamma^{(0)}_{ij,k}(\theta) - \tfrac{1}{2}T_{ijk}(\theta).$$

We derive a cleaner closed-form. Recall from Theorem 1.2.2 that $g_{ij}(\theta) = \mathbb{E}_\theta[\partial_i \ell_\theta\, \partial_j \ell_\theta] = -\mathbb{E}_\theta[\partial_i \partial_j \ell_\theta]$ (the Fisher–score-Hessian identity, valid under regularity conditions). An analogous calculation gives the skewness tensor in terms of mixed log-likelihood derivatives:

Under regularity conditions (interchange of differentiation and integration),

$$T_{ijk} = \mathbb{E}_\theta[\partial_i \ell\, \partial_j \ell\, \partial_k \ell] = \mathbb{E}_\theta[\partial_i \partial_j \ell\, \partial_k \ell] + \partial_k g^{(\mathrm{loc})}_{ij},$$

where the second expression uses the mixed-second-derivative score. (Here $g_{ij}^{(\mathrm{loc})} = \mathbb{E}_\theta[\partial_i \ell\, \partial_j \ell]$ denotes the Fisher information locally, but since $g_{ij} = g_{ij}^{(\mathrm{loc})}$ the superscript is only for notational clarity in the proof.) More usefully for computation, there is a direct third-derivative formula:

$$T_{ijk} = \partial_i \partial_j \partial_k \Phi(\theta),$$

valid for *exponential families* (where Φ is the cumulant generating function of Section 2.1).

Proof of the third-derivative formula for exponential families. For $p_\theta(x) = h(x)\exp(\theta^\top T(x) - \Phi(\theta))$, $\ell_\theta(x) = \log h(x) + \theta^\top T(x) - \Phi(\theta)$. So $\partial_i \ell_\theta = T_i(x) - \partial_i \Phi(\theta)$, and $\partial_i \ell_\theta(x) - \mathbb{E}_\theta[\partial_i \ell_\theta] = T_i(x) - \eta_i$ (using $\eta_i = \mathbb{E}_\theta[T_i] = \partial_i \Phi$). Thus:

$$T_{ijk} = \mathbb{E}_\theta[(T_i - \eta_i)(T_j - \eta_j)(T_k - \eta_k)] = \partial_i \partial_j \partial_k \Phi(\theta)$$

where the last equality is the standard cumulant identity: the r-th cumulant of a linear functional of the sufficient statistic equals the corresponding r-th partial derivative of Φ (proved by differentiating the moment generating function identity $\Phi(\theta) = \log \int h(x)\exp(\theta^\top T(x))\, dx$ three times). ■

Now compute the e-connection Christoffel symbols for exponential families. Recall $\Gamma_{ij,k}^{(0)} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij})$ and $g_{ij} = \partial_i \partial_j \Phi = T_{ij}^{(2)}$ (second derivatives of Φ, the Fisher metric of Theorem 2.2.1), so $\partial_k g_{ij} = \partial_k \partial_i \partial_j \Phi = T_{ijk}$. Hence:

$$\Gamma_{ij,k}^{(0)} = \tfrac{1}{2}(T_{jik} + T_{ijk} - T_{kij}) = \tfrac{1}{2} T_{ijk}$$

using total symmetry $T_{jik} = T_{ijk} = T_{kij}$. Therefore:

$$\Gamma_{ij,k}^{(e)} = \Gamma_{ij,k}^{(0)} - \tfrac{1}{2} T_{ijk} = \tfrac{1}{2} T_{ijk} - \tfrac{1}{2} T_{ijk} = 0.$$

For any exponential family, the e-connection Christoffel symbols vanish identically in the natural (θ) parametrisation:

$$\Gamma_{ij,k}^{(e)}(\theta) \equiv 0.$$

Consequently, $\nabla^{(e)}$ is *flat*, and the natural parameters $(\theta^1, \dots, \theta^n)$ are *affine coordinates* for $\nabla^{(e)}$. e-geodesics are straight lines in θ.

Geometric meaning. In the θ-coordinate, $\nabla^{(e)}$-parallel transport is trivial (the Christoffel symbols vanish, so the covariant derivative of a constant vector field along any curve is zero — "the basis vectors ∂_i don't tilt as you move through θ-space"). This means $\mathcal{S}$ with the e-connection looks, from the θ-chart, just like $\mathbb{R}^n$ with the flat Euclidean connection — the "curved" appearance of $\mathcal{S}$ in the ambient probability-simplex picture is entirely due to the curvature of the Fisher metric $g = \nabla^2 \Phi$, *not* to the e-connection. Flatness of $\nabla^{(e)}$ in θ is the reason that the maximum-likelihood estimator has such clean asymptotic geometry in natural parameters (Section 3.4 develops this).

The m-Connection and Mixture Flatness

m-Connection ($\alpha = -1$)

The *mixture connection* $\nabla^{(m)} \triangleq \nabla^{(-1)}$ has Christoffel symbols

$$\Gamma^{(m)}_{ij,k}(\theta) = \Gamma^{(0)}_{ij,k}(\theta) + \tfrac{1}{2}T_{ijk}(\theta).$$

By Theorem 3.1.5, $\nabla^{(m)}$ and $\nabla^{(e)}$ are dual with respect to $\mathbf{G}$. Flatness of $\nabla^{(m)}$ is established by switching to the expectation coordinates η.

In the expectation (η) parametrisation, the m-connection Christoffel symbols vanish:

$$\tilde{\Gamma}^{(m)}_{ij,k}(\eta) \equiv 0.$$

Consequently, $\nabla^{(m)}$ is flat, and the expectation parameters $(\eta_1, \dots, \eta_n)$ are affine coordinates for $\nabla^{(m)}$. m-geodesics are straight lines in η.

Proof. Under the coordinate change $\theta \to \eta$ (Jacobian $J_i{}^k = \partial\eta_i/\partial\theta^k = \partial_k\partial_i\Phi = g_{ki}$), the Christoffel symbols transform by the standard connection-transformation law. We use instead the cleaner direct approach: the m-connection is dual to the e-connection, and in a *dually flat* situation (established in Section 3.4: when one connection is flat in one coordinate and the dual is flat in the dual coordinate), the dual connection's Christoffel symbols in the dual coordinates are zero by definition of the Legendre duality structure. We verify this directly for the 1-dimensional case (Example 3.2.3 below) and quote the general result (Amari–Nagaoka 2000, Chapter 3). ■

Example — m-Flatness for Bernoulli: Verifying $\tilde{\Gamma}^{(m)}_{111} = 0$ in η-Coordinates

For the Bernoulli family in the $\eta = p$ coordinate (Example 3.1.6): we showed directly that $\Gamma^{(-1)}_{111}(\theta)\big|_{\theta=\mathrm{logit}(p)} = 0$ using $\Gamma^{(0)}_{111} = \frac{1}{2}g'(p)$ and $T_{111}(p) = (1-2p)g(p)^2 = -g'(p)$, so $\Gamma^{(-1)}_{111} = \frac{1}{2}g'(p) + \frac{1}{2}T_{111} = \frac{1}{2}g'(p) - \frac{1}{2}g'(p) = 0$. In the $\eta = p$ coordinate, the single m-connection Christoffel symbol vanishes identically for all $p \in (0,1)$, confirming Theorem 3.2.2 in dimension 1.
Geometric consequence: m-geodesics in the Bernoulli family are line segments in p-space, i.e. curves $p(t) = p_0 + t(p_1 - p_0)$ for $t \in [0,1]$. These are precisely *mixture paths* $p_t = (1-t)p_0 + tp_1$ — the connection is called "mixture" precisely because its geodesics are convex combinations (mixtures) of endpoint distributions.

Example — e-Flatness for the Gaussian Family in Natural Parameters

For $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\theta_1, \theta_2) = (\mu/\sigma^2, -1/(2\sigma^2))$ and $\Phi(\theta) = \mu^2/(2\sigma^2) + \frac{1}{2}\log\sigma^2$ (Section 2.5.2): $T_{ijk} = \partial_i\partial_j\partial_k\Phi$. Since Φ is a smooth function of θ, all its third partial derivatives are well-defined and in general nonzero — so $T_{ijk} \neq 0$ and $\Gamma^{(0)}_{ij,k} = \frac{1}{2}T_{ijk} \neq 0$ (the Levi-Civita connection is curved). But $\Gamma^{(e)}_{ij,k} = 0$ for all i, j, k (Theorem 3.2.1): the e-connection is flat in θ. Concretely, $T_{111} = \partial_1^3\Phi = \partial_1^3(\theta_1^2/(-4\theta_2) + \dots) = 0$ (only terms up to θ_1^2 appear in Φ, so the third θ_1-derivative vanishes), $T_{112} = \partial_1^2\partial_2\Phi = \partial_1^2(-1/(4\theta_2)) \dots$ actually since $\partial_1\Phi = -\theta_1/(2\theta_2)$, $\partial_1^2\Phi = -1/(2\theta_2) = \sigma^2$, and $\partial_1^2\partial_2\Phi = 1/(2\theta_2^2) = 2\sigma^4$.

And $T_{122} = \partial_1 \partial_2^2 \Phi$, $T_{222} = \partial_2^3 \Phi$ — all nonzero in general. Yet Theorem 3.2.1 guarantees $\Gamma^{(e)}_{ij,k} = \Gamma^{(0)}_{ij,k} - \frac{1}{2} T_{ijk} = \frac{1}{2} T_{ijk} - \frac{1}{2} T_{ijk} = 0$ regardless of these values. The e-flat structure of the Gaussian family is the geometric reason that exponential-family maximum-likelihood estimation has asymptotically linear (in θ) estimating equations and no correction terms for curvature.

The Meaning of e- and m-Geodesics

Since e-geodesics are lines in θ and m-geodesics are lines in η, and the Legendre map $\theta \leftrightarrow \eta$ is nonlinear (in general), e- and m-geodesics between the same two endpoints are *different curves* on $\mathcal{S}$. The e-geodesic through p_{θ_0} and p_{θ_1} is the one-parameter exponential family

$$p_{(1-t)\theta_0 + t\theta_1}(x) \;\propto\; p_{\theta_0}(x)^{1-t}\, p_{\theta_1}(x)^t, \qquad t \in [0,1].$$

(A product-of-powers: the geometric mixture, or *exponential arc*.)

The m-geodesic through p_0 and p_1 is the *arithmetic mixture*:

$$p_t = (1-t)\, p_{\eta_0} + t\, p_{\eta_1}, \qquad t \in [0,1],$$

i.e. the straight line in the η-simplex, convex combination of the two densities (pointwise). The m-geodesic stays inside the space of probability distributions ($\sum_x p_t(x) = 1$ at every t), but it need not stay inside the exponential family $\mathcal{S}$ unless $\mathcal{S}$ is a *mixture family* (a set closed under convex combinations, exactly the m-flat families of Theorem 3.2.2).

Duality of e- and m-geodesics. e-geodesics are exponential arcs (geometric mixtures, straight in θ, used in robust statistics and generalised exponential tilting); m-geodesics are arithmetic mixtures (convex combinations, straight in η, used in Bayesian model averaging and mixture modelling). The two are equal only when the Fisher metric is flat (e.g. for the Gaussian family restricted to fixed variance, which is both an exponential and a mixture family simultaneously in the 1-dimensional normal mean submodel).

Summary of Section 3.2. For exponential families: (i) $\Gamma^{(e)}_{ij,k} \equiv 0$ in θ: the e-connection is flat, natural parameters are e-affine, e-geodesics are lines $\theta(t) = (1-t)\theta_0 + t\theta_1$ (exponential arcs). (ii) $\tilde{\Gamma}^{(m)}_{ij,k} \equiv 0$ in η: the m-connection is flat, expectation parameters are m-affine, m-geodesics are lines $\eta(t) = (1-t)\eta_0 + t\eta_1$ (arithmetic mixtures). (iii) The two flatnesses are dual to each other via the Fisher metric (Theorem 3.1.5) and the Legendre map $\theta \leftrightarrow \eta$. This dual flatness — one space, two flat coordinate systems, related by a nonlinear Legendre transform and tied together by the Fisher metric — is the content of the fundamental theorem of dually flat spaces in Section 3.4.

3.3 The Fundamental Theorem of Dually Flat Spaces

Dually Flat Manifolds: Definition and First Properties

Dually Flat Manifold

A Riemannian manifold $(\mathcal{M}, g)$ equipped with a pair of torsion-free connections (∇, ∇^*) is *dually flat* if:

1. ∇ is flat (zero curvature) with affine coordinates $\theta = (\theta^i)$, and
2. ∇^* is flat with affine coordinates $\eta = (\eta_i)$, and
3. ∇, ∇^* are dual with respect to g: $Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z)$.

Sections 3.1–3.2 established that every exponential family satisfies (1)–(3) with $\nabla = \nabla^{(e)}$, $\nabla^* = \nabla^{(m)}$, $g = \mathbf{G}$, $\theta =$ natural parameters, $\eta =$ expectation parameters. What does the dually flat structure *enforce* on the relationships among θ, η, g, and the divergence?

Let $(\mathcal{M}, g, \nabla, \nabla^*)$ be dually flat with ∇-affine coordinates θ and ∇^*-affine coordinates η. Then there exist strictly convex potential functions $\Phi : \mathbb{R}^n \to \mathbb{R}$ and $\Psi : \mathbb{R}^n \to \mathbb{R}$ such that:

1. $g_{ij}(\theta) = \partial_i \partial_j \Phi(\theta)$: the metric is the Hessian of Φ in θ.
2. $g^{ij}(\eta) = \partial^i \partial^j \Psi(\eta)$: the inverse metric is the Hessian of Ψ in η.
3. $\eta_i = \partial_i \Phi(\theta)$ and $\theta^i = \partial^i \Psi(\eta)$: the coordinate systems are related by the Legendre transform.
4. $\Phi(\theta) + \Psi(\eta) = \theta^\top \eta$ (the Young–Fenchel identity).

Proof sketch. (1): In θ-coordinates, ∇ is flat so $\Gamma^k_{ij} = 0$; the duality condition $\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}$ with $\Gamma_{ki,j} = 0$ (since ∇ is flat in θ, $\Gamma^{(e)}_{ki,j} = 0$, so $g(\nabla_{\partial_k}\partial_i, \partial_j) = 0$) gives $\partial_k g_{ij} = \Gamma^*_{kj,i}$. This, plus g_{ij} positive definite and symmetric, implies $g_{ij} = \partial_i \partial_j \Phi$ for some strictly convex Φ (integrability follows from symmetry of g_{ij}; strict convexity from positive definiteness). (3) and (4) follow from the standard Legendre-transform construction (Section 2.2): given Φ strictly convex, define $\eta_i = \partial_i \Phi$ and $\Psi(\eta) = \sup_\theta \{\theta^\top \eta - \Phi(\theta)\}$, giving the Young–Fenchel identity and the inverse relationship $\theta^i = \partial^i \Psi$. (2) is then the inverse-Hessian relation of Theorem 2.2.3. ■

The Canonical Divergence of a Dually Flat Space

Canonical Divergence

On a dually flat manifold $(\mathcal{M}, g, \nabla, \nabla^*)$ with potentials Φ, Ψ and coordinate pairs (θ, η), the *canonical divergence* from P (with coordinates θ_P, η_P) to Q (with θ_Q, η_Q) is

$$D(P\|Q) \;\triangleq\; \Phi(\theta_P) + \Psi(\eta_Q) - \theta_P^\top \eta_Q \;=\; B_\Phi(\theta_P : \theta_Q) \;=\; B_\Psi(\eta_P : \eta_Q)\,.$$

The three expressions are equal:

$$\begin{aligned} B_\Phi(\theta_P : \theta_Q) &= \Phi(\theta_P) - \Phi(\theta_Q) - \nabla\Phi(\theta_Q)^\top(\theta_P - \theta_Q) \\ &= \Phi(\theta_P) - \Phi(\theta_Q) - \eta_Q^\top(\theta_P - \theta_Q) \\ &= \Phi(\theta_P) + \Psi(\eta_Q) - \theta_Q^\top \eta_Q - \eta_Q^\top \theta_P + \eta_Q^\top \theta_Q \\ &= \Phi(\theta_P) + \Psi(\eta_Q) - \theta_P^\top \eta_Q, \end{aligned}$$

using the Young–Fenchel identity $\Phi(\theta_Q) + \Psi(\eta_Q) = \theta_Q^\top \eta_Q$. The equality $B_\Phi(\theta_P : \theta_Q) = B_\Psi(\eta_P : \eta_Q)$ was proved in Proposition 2.3.3. Combined with Theorem 2.3.2 (KL is the Bregman divergence of Φ), we recover:

For exponential families with the dually flat structure $(\nabla^{(e)}, \nabla^{(m)}, \mathbf{G})$:

$$D_{\mathrm{KL}}\big(p_{\theta_P} \,\|\, p_{\theta_Q}\big) = D(P\|Q) = \Phi(\theta_P) + \Psi(\eta_Q) - \theta_P^\top \eta_Q.$$

This is a far-reaching statement: the KL divergence is not merely a convenient measure of distributional discrepancy, but the *canonical* (i.e. unique, up to scale, invariant under the symmetry group of the dually flat structure) divergence of the information-geometric manifold $\mathcal{S}$. Among all divergence functions one could write down, KL is singled out by the dual structure, just as the squared Euclidean distance is singled out by Euclidean flat geometry.

The Pythagorean Theorem in Dually Flat Geometry

The Pythagorean theorem of Section 2.4 can now be restated in purely geometric terms, without reference to Bregman divergences or explicit coordinate expressions.

Let $(\mathcal{M}, g, \nabla, \nabla^*)$ be dually flat, with canonical divergence D. Let $P, Q, R \in \mathcal{M}$. Then

$$D(P\|R) = D(P\|Q) + D(Q\|R)$$

if and only if the ∇^*-geodesic from P to Q and the ∇-geodesic from Q to R are *orthogonal at* Q with respect to g: $g_Q(\dot\gamma^*_{PQ}(1), \dot\gamma_{QR}(0)) = 0$.

This is the geometric version of Theorem 2.4.2 (Bregman Pythagorean theorem for affine C): "the ∇^*-geodesic from P to its ∇-projection Q^* onto a ∇-flat submanifold $\mathcal{E}$ is g-orthogonal to $\mathcal{E}$" is the geometric content of "the tangent to the ∇-geodesic through Q^* lies in the tangent space to $\mathcal{E}$, and is g-orthogonal to the ∇^*-geodesic from P to Q^*".

For the statistical manifold: $\mathcal{E}$ is an exponential subfamily ($\nabla^{(e)}$-flat, geodesics are lines in θ), the I-projection $Q^* = \arg\min_{R\in\mathcal{E}} D_{\mathrm{KL}}(P \,\|\, P_R)$ is the $\nabla^{(e)}$-projection of P onto $\mathcal{E}$, and g-orthogonality of the $\nabla^{(m)}$-geodesic from P to Q^* and $\mathcal{E}$ at Q^* is the geometric statement of Theorem 2.4.4. Similarly, $\mathcal{M}$ a mixture family ($\nabla^{(m)}$-flat) and Q^* the M-projection gives Theorem 2.4.3. The two cases are dual to each other, exchanging $\nabla^{(e)} \leftrightarrow \nabla^{(m)}$ and $\theta \leftrightarrow \eta$.

Three-point identity and KL decomposition. An important corollary of the Pythagorean theorem is the *three-point identity* (also called the *generalised law of cosines* for dually flat spaces). For any P, Q, R:

$$D(P\|R) = D(P\|Q) + D(Q\|R) - (\theta_P - \theta_Q)^\top (\eta_R - \eta_Q),$$

where (θ, η) are the dual coordinates. The cross-term $(\theta_P - \theta_Q)^\top(\eta_R - \eta_Q)$ vanishes exactly when the m-geodesic PR and e-geodesic QR are g-orthogonal at R, or equivalently when the "inner product" in the dual coordinates is zero. This identity is the information-geometric analogue of the law of cosines $\|PR\|^2 = \|PQ\|^2 + \|QR\|^2 - 2\|PQ\|\|QR\|\cos\angle$, with the dual-coordinate inner product playing the role of $\cos\angle$.

Example — Variational Inference as m-Projection and the ELBO

In variational Bayesian inference, one approximates an intractable posterior $p(\theta|x)$ by the distribution q_ϕ in a variational family $\mathcal{Q} = \{q_\phi : \phi \in \Phi\}$ that minimises $D_{\mathrm{KL}}(q_\phi \,\|\, p(\theta|x))$ — the m-projection of the posterior onto $\mathcal{Q}$. The variational objective is the *evidence lower bound* (ELBO):

$$\mathrm{ELBO}(\phi) = \mathbb{E}_{q_\phi}[\log p(x,\theta) - \log q_\phi(\theta)] = \log p(x) - D_{\mathrm{KL}}(q_\phi \,\|\, p(\theta|x)) \,.$$

Maximising the ELBO over ϕ is exactly minimising $D_{\mathrm{KL}}(q_\phi \,\|\, p(\theta|x))$, i.e. finding the m-projection $q^* = \pi_{\mathcal{Q}}^{(m)}(p(\cdot|x))$.
When $\mathcal{Q}$ is the mean-field family (product distributions $q_\phi(\theta) = \prod_k q_k(\theta_k)$ over blocks), the m-projection decomposes into coordinate-wise updates: each factor q_k^* is the m-projection of $p(\theta|x)$ onto the k-th marginal constraint, and the coordinate ascent variational inference (CAVI) algorithm alternates these projections — exactly the alternating m-projection algorithm of Section 2.4.
The Pythagorean decomposition gives:

$$D_{\mathrm{KL}}(q \,\|\, p(\theta|x)) = D_{\mathrm{KL}}(q \,\|\, q^*) + D_{\mathrm{KL}}(q^* \,\|\, p(\theta|x)) \,, \quad \forall q \in \mathcal{Q},$$

so $D_{\mathrm{KL}}(q^* \,\|\, p(\theta|x))$ is the irreducible approximation error (the "approximation gap" from the best possible fit in $\mathcal{Q}$), while $D_{\mathrm{KL}}(q \,\|\, q^*)$ is the optimisation gap (how far the current iterate q is from the variational optimum q^*). This decomposition, a direct application of Theorem 3.3.3, is the foundation of the recent "tighter ELBO" and "importance-weighted" variational inference literature.

Example — e-Projection and Maximum Entropy

By duality, the e-projection of a distribution P onto a mixture-flat set $\mathcal{M}$ (minimising $D_{\mathrm{KL}}(\mathcal{M} \,\|\, P)$, i.e. $\arg\min_{Q \in \mathcal{M}} D_{\mathrm{KL}}(Q \,\|\, P)$) is the maximum entropy distribution subject to the moment constraints that define $\mathcal{M}$. Concretely: if $\mathcal{M} = \{Q : \mathbb{E}_Q[f_k] = \mu_k, k = 1, \dots, K\}$ (an m-flat constraint set), then the e-projection $Q^* = \pi_{\mathcal{M}}^{(e)}(P)$ satisfies $Q^*(x) \propto P(x) \exp(\sum_k \lambda_k^* f_k(x))$ (an exponential tilt of P) and is the *maximum relative entropy* distribution subject to the constraints — the information-geometric restatement of the principle of maximum entropy (Jaynes 1957).
The Pythagorean theorem gives the decomposition:

$$D_{\mathrm{KL}}(Q \,\|\, P) = D_{\mathrm{KL}}(Q \,\|\, Q^*) + D_{\mathrm{KL}}(Q^* \,\|\, P) \,, \quad \forall Q \in \mathcal{M},$$

where $D_{\mathrm{KL}}(Q^* \,\|\, P)$ is the minimum KL achieved by any distribution in $\mathcal{M}$, and $D_{\mathrm{KL}}(Q \,\|\, Q^*) \geq 0$ measures the additional divergence of Q from Q^* within $\mathcal{M}$.

Curvature of the e- and m-Connections vs. Levi-Civita

Section 3.2 showed $\nabla^{(e)}$ and $\nabla^{(m)}$ are individually flat. How does this compare with $\nabla^{(0)}$ (Levi-Civita)?

For a statistical manifold $(\mathcal{S}, \mathbf{G}, \nabla^{(\alpha)})$ with skewness tensor T_{ijk}:

1. $R^{(e)} = R^{(m)} = 0$ for exponential families (Theorems 3.2.1, 3.2.2).
2. $R^{(0)} = 0$ iff $T_{ijk} = 0$ for all i, j, k (the skewness tensor vanishes identically).
3. More generally, for the α-connection:

$$R^{(\alpha)}_{kl,ij} = R^{(0)}_{kl,ij} - \alpha\, \nabla^{(0)}_{[k} T_{l]ij} + \frac{\alpha^2}{4}(T_{kia} g^{ab} T_{ljb} - T_{kja} g^{ab} T_{lib}),$$

   where $T_{kia} g^{ab} T_{ljb} \triangleq \sum_{a,b} T_{kia} (\mathbf{G}^{-1})^{ab} T_{ljb}$ is the "square" of T contracted with the inverse metric.
4. For exponential families: $R^{(0)}_{kl,ij} = \frac{\alpha^2-1}{4}(T_{kia} g^{ab} T_{ljb} - T_{kja} g^{ab} T_{lib})$ (setting $R^{(\pm 1)} = 0$ in the formula above), so

$$R^{(0)}_{kl,ij} = -\frac{1}{4}(T_{kia} g^{ab} T_{ljb} - T_{kja} g^{ab} T_{lib}).$$

Result (4) says: for exponential families, the Levi-Civita curvature is nonzero whenever the skewness tensor T_{ijk} is nonzero (i.e. whenever the family is not "trivially Gaussian" — zero third cumulant means Gaussian, by the Marcinkiewicz theorem). So the Gaussian family ($T \equiv 0$ in the natural parameters, since Φ is quadratic and thus $T_{ijk} = \partial_i \partial_j \partial_k \Phi = 0$) has $R^{(0)} = 0$ as well (the Fisher-Rao geometry of the Gaussian family is flat!), while the Bernoulli, Poisson, and other exponential families have $R^{(0)} \neq 0$ — they are curved in the Levi-Civita sense, even though $\nabla^{(\pm 1)}$ are flat.

Example — Gaussian Family: Globally Flat (All Three Connections)

For $\mathcal{N}(\mu, \sigma^2)$ with natural parameters $\theta = (\mu/\sigma^2, -1/(2\sigma^2))$: $\Phi(\theta) = -\theta_1^2/(4\theta_2) - \frac{1}{2}\log(-2\theta_2)$ is quadratic in θ_1 and a function of θ_2 alone (not quadratic in θ_2, but differentiating three times: $T_{111} = \partial_1^3 \Phi = 0$, $T_{112} = \partial_1^2 \partial_2 \Phi = \partial_2[-1/(2\theta_2)] = 1/(2\theta_2^2)$, $T_{122} = \partial_1 \partial_2^2 \Phi = 0$ (since $\partial_1 \Phi = -\theta_1/(2\theta_2)$ and $\partial_2[-\theta_1/(2\theta_2)] = \theta_1/(2\theta_2^2)$, whose θ_1-derivative is $1/(2\theta_2^2)$... let me recount: $\partial_1 \Phi = -\theta_1/(2\theta_2)$; $\partial_2 \partial_1 \Phi = \theta_1/(2\theta_2^2)$; $\partial_1^2 \partial_2 \Phi = 1/(2\theta_2^2) = T_{112} \neq 0$. So T is not identically zero for the Gaussian family — only $T_{111} = T_{222}(\sigma) = \partial_2^3 \Phi = \partial_2[1/(2\theta_2^2) - ...]$ needs checking: $\partial_2^2 \Phi = \partial_2[-1/(2\theta_2)] = 1/(2\theta_2^2)$; ... actually: $\Phi(\theta) = -\theta_1^2/(4\theta_2) - \frac{1}{2}\log(-2\theta_2)$; so $\partial_2 \Phi = \theta_1^2/(4\theta_2^2) - \frac{1}{2} \cdot (-2)/(-2\theta_2) = \theta_1^2/(4\theta_2^2) - 1/(2\theta_2)$; $\partial_2^2 \Phi = -\theta_1^2/(2\theta_2^3) + 1/(2\theta_2^2)$; $\partial_2^3 \Phi = 3\theta_1^2/(2\theta_2^4) - 1/\theta_2^3 = T_{222} \neq 0$. So the Gaussian family has nonzero T and thus $R^{(0)} \neq 0$ in general — contradicting the claim above. The correct statement is: the *one-dimensional Gaussian family with fixed variance* (the normal-mean model: only μ varies, $\sigma^2 = 1$ fixed) has $\theta = \mu$, $\Phi(\theta) = \theta^2/2$ (quadratic!), so $T_{111} = 0$ and $R^{(0)} = 0$. The full two-parameter Gaussian family has nonzero $R^{(0)}$; the Gaussian family's distinctiveness lies in the *natural*-parameter curvature $R^{(0)}$, not its being zero (it is not, in general). The flat cases are: (a) normal mean model (fixed variance), (b) full Gaussian family in a specific dimension-dependent geometry. See Amari–Nagaoka Chapter 4 for the classification of dually flat families beyond exponential families.

Summary of Section 3.3. A dually flat manifold $(\mathcal{M}, g, \nabla, \nabla^*)$ admits dual potential functions (Φ, Ψ) related by Legendre transform and dual affine coordinates (θ, η) related by $\eta = \nabla\Phi$ (Theorem 3.3.1). The canonical divergence $D(P\|Q) = \Phi(\theta_P) + \Psi(\eta_Q) - \theta_P^\top \eta_Q$ equals the Bregman divergence of Φ (Definition 3.3.2), and for the statistical manifold equals KL (Theorem 3.3.3). The Pythagorean theorem $D(P\|R) = D(P\|Q) + D(Q\|R)$ holds exactly iff the ∇^*-geodesic from P to Q is g-orthogonal to the ∇-geodesic from Q to R (Theorem 3.3.4) — the geometric unification of all the I- and M-projection results of Part 2. Levi-Civita curvature $R^{(0)} \neq 0$ in general (Proposition 3.3.5) even when $R^{(\pm 1)} = 0$: dually flat $\neq$ Riemannian flat, except for the normal-mean model.

Eguchi's Theorem: Contrast Functions and the Recovery of Dual Geometry

The canonical divergence of a dually flat manifold is a special case of a much more general construction. Eguchi (1983, 1992) showed that *any* smooth, non-degenerate contrast function (a generalised divergence) on a manifold $\mathcal{M}$ induces, via its second-order derivatives, a Riemannian metric and a pair of dual torsion-free connections on $\mathcal{M}$. This is a converse to the duality theory: rather than starting with a geometry and deriving a divergence, one starts with a divergence and recovers the geometry.

Contrast Function (Eguchi)

A smooth function $D : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ is a *contrast function* if:

1. $D(P, Q) \geq 0$ for all $P, Q \in \mathcal{M}$, with equality iff $P = Q$.
2. D is smooth in a neighbourhood of the diagonal $\{(P, P) : P \in \mathcal{M}\}$.
3. All first-order partial derivatives of $D(p_\theta, p_\phi)$ vanish at $\theta = \phi$: $(\partial/\partial\theta^i) D(p_\theta, p_\phi)\big|_{\theta=\phi} = 0$ for all i (and similarly in the second argument).

Let $D : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ be a contrast function on a statistical manifold $\mathcal{S}$. Define, in local coordinates θ (writing $\partial_i = \partial/\partial\theta^i$ and $\partial_i' = \partial/\partial\phi^i$ for derivatives with respect to the first and second arguments respectively):

$$
\begin{aligned}
g_{ij}(\theta) &\triangleq -\partial_i \partial_j' \, D(p_\theta, p_\phi)\big|_{\theta=\phi}, \\
\Gamma_{ij,k}(\theta) &\triangleq -\partial_i \partial_j \partial_k' \, D(p_\theta, p_\phi)\big|_{\theta=\phi}, \\
\Gamma^*_{ij,k}(\theta) &\triangleq -\partial_i' \partial_j' \partial_k \, D(p_\theta, p_\phi)\big|_{\theta=\phi}.
\end{aligned}
$$

Then g_{ij} is a positive-definite Riemannian metric on $\mathcal{S}$, and Γ, Γ^* are Christoffel symbols (with respect to g) of a pair of dual torsion-free connections ∇, ∇^* satisfying the duality identity $Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z)$ for all vector fields X, Y, Z.

The elegance of Eguchi's theorem is that the choice of contrast function is equivalent to the choice of geometry: every geometry (metric + dual connections) arises from some contrast function, and different contrast functions inducing the same metric g but different connections correspond to the same Riemannian geometry but different (dual) statistical structures.

Example — Eguchi Reconstruction of the Fisher Metric from KL

For the KL divergence $D_{\mathrm{KL}}(p_\theta \,\|\, p_\phi) = \int p_\theta \log(p_\theta/p_\phi)\, d\mu$: the mixed second derivative at $\theta = \phi$ is

$$-\partial_i \partial'_j D_{\mathrm{KL}}(p_\theta \,\|\, p_\phi)\,\big|_{\theta=\phi} = \mathbb{E}_\theta[\partial_i \ell_\theta(X)\, \partial_j \ell_\theta(X)] = g^{\mathrm{Fisher}}_{ij}(\theta),$$

recovering the Fisher metric exactly (the cross-term $\partial'_j(\log p_\phi)$ evaluated at $\phi = \theta$ gives $-\partial_j \ell_\theta$, and the expectation of a product of scores gives the Fisher matrix entry). The third-order derivative $-\partial_i \partial_j \partial'_k D_{\mathrm{KL}}(p_\theta \,\|\, p_\phi)\,\big|_{\theta=\phi}$ gives the Christoffel symbols $\Gamma^{(e)}_{ij,k}$ of the exponential connection (this requires a computation involving $\partial_k(\log p_\phi)$ at $\phi = \theta$ and the score moment $\mathbb{E}_\theta[\partial_i \ell\, \partial_j \ell\, \partial_k \ell]$), while $-\partial'_i \partial'_j \partial_k D_{\mathrm{KL}}(p_\theta \,\|\, p_\phi)\,\big|_{\theta=\phi}$ gives $\Gamma^{(m)}_{ij,k}$, confirming the duality $(\nabla^{(e)})^* = \nabla^{(m)}$ in the sense of Eguchi's theorem.

Second-order expansion of a contrast function. Near the diagonal, any contrast function admits the Taylor expansion

$$D(p_\theta,\, p_{\theta+d\theta}) = \tfrac{1}{2} \sum_{i,j} g_{ij}(\theta)\, d\theta^i\, d\theta^j + \tfrac{1}{6} \sum_{i,j,k} (2\Gamma_{ij,k}(\theta) - \Gamma^*_{ij,k}(\theta))\, d\theta^i\, d\theta^j\, d\theta^k + O(\|d\theta\|^4).$$

The leading term is the Riemannian squared distance $\frac{1}{2} g(d\theta, d\theta)$, which is the same for all contrast functions inducing the same metric g. The third-order term encodes the *asymmetry* of the divergence (since $D(P,Q) \neq D(Q,P)$ in general), governed by $2\Gamma - \Gamma^*$. For a symmetrised divergence $D_{\mathrm{sym}}(P,Q) = D(P,Q) + D(Q,P)$, the third-order term vanishes (since swapping (P,Q) swaps Γ and Γ^*, so the sum gives $2\Gamma - \Gamma^* + 2\Gamma^* - \Gamma = \Gamma + \Gamma^*$ minus twice the symmetric part — in fact the full third-order coefficient is $(\Gamma - \Gamma^*)/2$ for the symmetrised divergence, which vanishes iff $\Gamma = \Gamma^*$, i.e. the connection is self-dual, i.e. equal to Levi-Civita $\nabla^{(0)}$). The fourth-order term, which governs the rate at which the symmetrised divergence grows along a geodesic, encodes the α-connection curvature $R^{(0)}$ — this is the information-geometric meaning of Riemannian curvature.

The significance of Eguchi's theorem for this volume is twofold. First, it gives a completely coordinate-free, divergence-first derivation of dual geometry, useful when exponential family structure is unavailable. Second, it establishes that the information-geometric framework extends far beyond exponential families: f-divergences (Appendix F), β-divergences, maximum mean discrepancy, and other dissimilarity measures each give rise to a legitimate information geometry with their own Fisher-like metric and dual connections. The particular geometry of exponential families is singled out by the additional requirement of *invariance under sufficient statistics* (Chentsov's theorem, Section 3.1.2), which forces the metric to be the Fisher metric and the connection family to be the α-connections.

3.4 Asymptotic Statistics and the Geometry of Maximum-Likelihood Estimation

The dually flat structure of Sections 3.1–3.3 has immediate, concrete consequences for the geometry of maximum-likelihood estimation (MLE). This section develops those consequences, making precise the senses in which (i) the MLE is the e-projection of the empirical distribution onto $\mathcal{S}$, (ii) the Fisher information matrix governs estimation precision, and (iii) the α-connection curvature appears as a higher-order correction term in the asymptotic expansion of the bias of the MLE.

The MLE as an e-Projection

Given n i.i.d. observations $x_1, \dots, x_n$ from an unknown $p_{\theta^*} \in \mathcal{S}$, the MLE $\hat{\theta}_n$ maximises the log-likelihood $\sum_{i=1}^n \ell_\theta(x_i) = n\mathbb{E}_{\hat{P}_n}[\ell_\theta]$, where $\hat{P}_n$ is the empirical distribution. Equivalently, $\hat{\theta}_n$ minimises the KL divergence from $\hat{P}_n$ to $\mathcal{S}$:

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} D_{\mathrm{KL}}\big(\hat{P}_n \,\|\, p_\theta\big)\,.$$

By Theorem 3.3.3 (KL = canonical divergence) and Theorem 2.4.4 (I-projection onto θ-affine $\mathcal{E}$), this is the *e-projection* (I-projection) of the empirical distribution $\hat{P}_n$ onto the exponential family $\mathcal{S}$. Since $\mathcal{S}$ is e-flat (Theorem 3.2.1), the e-projection is characterised by a clean geometric condition: the m-geodesic from $\hat{P}_n$ to $\hat{\theta}_n$ is **G**-orthogonal to $\mathcal{S}$ at $\hat{\theta}_n$.

$\hat{\theta}_n$ is the MLE iff

$$\eta(\hat{\theta}_n) = \hat{\eta}_n \triangleq \frac{1}{n}\sum_{i=1}^n T(x_i),$$

i.e. the expectation parameters at the MLE equal the sample averages of the sufficient statistics. Equivalently, the moment-matching condition $\mathbb{E}_{\hat{\theta}_n}[T(X)] = \hat{\eta}_n$ holds at the MLE.

Proof. The first-order condition $\partial_j D_{\mathrm{KL}}\big(\hat{P}_n \,\|\, p_\theta\big) = 0$ is $\mathbb{E}_{\hat{P}_n}[\partial_j \ell_\theta(X)] = 0$, i.e. $\sum_{i=1}^n (T_j(x_i) - \partial_j\Phi(\theta))/n = 0$, giving $\partial_j\Phi(\hat{\theta}_n) = \hat{\eta}_{n,j}$ for all j. Since $\nabla\Phi(\theta) = \eta(\theta)$ (Theorem 2.1.4), this is $\eta(\hat{\theta}_n) = \hat{\eta}_n$. ■

Geometric reading. The m-geodesic in η-coordinates from $\hat{P}_n$ (with expectation parameter $\hat{\eta}_n$, even if $\hat{P}_n$ is not in $\mathcal{S}$) to its e-projection $\hat{\theta}_n$ is the straight line $\eta(t) = \hat{\eta}_n$ — a degenerate geodesic, because the moment-matching condition $\eta(\hat{\theta}_n) = \hat{\eta}_n$ means $\hat{P}_n$ and $\hat{\theta}_n$ *have the same η-coordinate*. The e-projection does not change the expectation parameter — it finds the member of $\mathcal{S}$ with the same sufficient-statistic moments as the empirical distribution. This is the information-geometric explanation of the fundamental fact that the MLE of an exponential family is the method-of-moments estimator.

The Cramér–Rao Bound Revisited: Geometric Proof

The Cramér–Rao bound (Section 1.3) states that no unbiased estimator $\hat{\theta}$ of θ can have variance matrix smaller than $\mathbf{G}(\theta)^{-1}$ (in the positive-semidefinite sense). The dually flat geometry provides a clean geometric proof.

For an unbiased estimator $\hat{\theta}$, the estimating function $\psi_\theta(x) = \hat{\theta}(x) - \theta$ satisfies $\mathbb{E}_\theta[\psi_\theta(X)] = 0$ and $\mathbb{E}_\theta[\partial_j \psi^i_\theta(X)] \equiv -\delta^i_j$ (by differentiation under the integral). Writing $S_j(x) = \partial_j \ell_\theta(x)$ (the score), the covariance inequality gives, for any vectors $a, b \in \mathbb{R}^n$,

$$\mathrm{Cov}_\theta(a^\top\hat{\theta}, b^\top S) = a^\top \mathrm{Cov}_\theta(\hat{\theta}, S)\, b = -a^\top b \quad \Rightarrow \quad \mathrm{Var}_\theta(a^\top\hat{\theta})\, b^\top \mathbf{G}(\theta) b \geq (a^\top b)^2,$$

by Cauchy–Schwarz. Setting $b = \mathbf{G}(\theta)^{-1}a$: $\mathrm{Var}_\theta(a^\top\hat{\theta}) \geq a^\top\mathbf{G}(\theta)^{-1}a$ for all a — exactly the matrix-inequality form of the Cramér–Rao bound. The geometric content: $\mathrm{Cov}_\theta(\hat{\theta}, S) = -\mathrm{Id}$

says the estimator $\hat{\theta}$ and the score S are "anti-correlated in a canonical way", and this anti-correlation is determined by the Fisher metric — the Cramér–Rao bound is the Cauchy–Schwarz inequality applied to the inner product induced by $\mathbf{G}^{-1}$ on the space of estimators.

Example — Cramér–Rao for the Gaussian Mean

For $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ (known variance), the MLE is $\bar{X}_n = n^{-1}\sum X_i$ with $\mathrm{Var}_\mu(\bar{X}_n) = 1/n$. The Fisher information (one observation) is $\mathbf{G}(\mu) = 1$, so the Cramér–Rao bound gives $\mathrm{Var}_\mu(\hat{\mu}) \geq 1/n$ for any unbiased $\hat{\mu}$. The MLE $\bar{X}_n$ achieves this: it is *efficient* (attains the bound). In the language of Section 3.4.1: $\hat{\mu}_n = \bar{X}_n$ is exactly the sample mean of the sufficient statistic $T(x) = x$, which equals the MLE by Proposition 3.4.1, and it achieves the Cramér–Rao bound because the Gaussian mean model is an exponential family (hence e-flat) and $\bar{X}_n$ is the e-projection of $\hat{P}_n$ onto $\mathcal{S}$.

Higher-Order Asymptotics and α-Connection Curvature

Proposition 3.4.1 established the first-order efficiency of the MLE via the e-flat geometry. Higher-order asymptotics ($O(1/n)$ bias terms, $O(1/n^2)$ variance corrections) involve the *curvature* of $\nabla^{(0)}$ (the Levi-Civita/Fisher-Rao curvature, which is nonzero for non-Gaussian exponential families, Proposition 3.3.5). Two results illustrate this.

Bartlett correction. The log-likelihood ratio statistic $\Lambda_n = 2\log(p_{\hat{\theta}_n}/p_{\theta^*})$ is asymptotically χ^2_k ($k =$ dimension of $\mathcal{S}$), with an $O(1/n)$ correction term (the *Bartlett correction factor* $c = 1 + b/n$) that depends on the second-order cumulants of the log-likelihood: $n\Lambda_n/c \to \chi^2_k$ more accurately. The correction factor b is given by a formula involving T_{ijk} (the skewness tensor) contracted with $\mathbf{G}^{-1}$:

$$b = \frac{1}{6}\sum_{i,j,k,l,m,n} T_{ijk}\, g^{il}\, g^{jm}\, g^{kn}\, T_{lmn} - \frac{1}{4}\sum_{i,j,k,l} T_{ijk}\, g^{jl}\, T^{(4)}_{ikl},$$

where $T^{(4)}_{ikl}$ involves fourth cumulants of the score — the curvature of $\nabla^{(0)}$ appears explicitly via the $T_{ijk}T_{lmn}$-type contractions.

Amari's α-connection bias formula. For an estimator $\hat{\theta}$ based on a score equation $\sum \psi_\theta(x_i) = 0$, the $O(1/n)$ bias of $\hat{\theta}$ in the direction of θ^* is, to leading order,

$$\mathbb{E}_{\theta^*}[\hat{\theta}^k - \theta^{*k}] = -\frac{1}{2n}\sum_{i,j} g^{ki}\, \tilde{H}_{ij} + O(n^{-2}),$$

where $\tilde{H}_{ij}$ is a tensor built from $\Gamma^{(1-\alpha)}_{ij,k}$ and $\mathbf{G}^{-1}$, and α is the *order* of the estimating function ψ. For the MLE ($\psi = \partial_j \ell_\theta$), $\alpha = 1$ (the MLE is a "e-estimating function"), and the leading α-connection Christoffel symbols are those of $\nabla^{(0)}$ ($\alpha = 1$: $1 - \alpha = 0$). This makes the MLE bias formula second-order in the e-connection curvature, consistent with Proposition 3.4.1 showing that the MLE is first-order efficient (zero leading bias).

Natural Gradient Descent

The information-geometric perspective immediately suggests a modification of gradient descent for statistical learning: instead of moving in the direction of the steepest gradient in *Euclidean* parameter space (i.e. $\theta_{t+1} = \theta_t - \varepsilon \nabla_\theta L$, which treats all directions as equal regardless of the Fisher metric), move in the steepest direction in the *Riemannian* ($\mathbf{G}$-weighted) sense.

Natural Gradient

The *natural gradient* of a loss $L(\theta)$ is

$$\tilde{\nabla}_\theta L \triangleq \mathbf{G}(\theta)^{-1}\nabla_\theta L.$$

The natural gradient descent update is $\theta_{t+1} = \theta_t - \varepsilon\,\mathbf{G}(\theta_t)^{-1}\nabla_\theta L(\theta_t)$.

The natural gradient direction $\mathbf{G}(\theta)^{-1}\nabla_\theta L(\theta)$ is the unique direction of steepest descent of L on the Riemannian manifold $(\mathcal{S}, \mathbf{G})$: for a unit-speed perturbation $\theta \to \theta + d\theta$ with $d\theta^\top \mathbf{G}(\theta)\, d\theta = \varepsilon^2$, the directional derivative $\nabla_\theta L^\top d\theta$ is maximised (most negative) when $d\theta \propto -\mathbf{G}(\theta)^{-1}\nabla_\theta L$.

Proof. By Cauchy–Schwarz for the $\mathbf{G}$-inner product: $|\nabla_\theta L^\top d\theta| = |(d\theta)^\top \mathbf{G}^{1/2}\mathbf{G}^{-1/2}\nabla_\theta L| \le \|d\theta\|_{\mathbf{G}} \cdot \left\|\mathbf{G}^{-1/2}\nabla_\theta L\right\|_{\text{Eucl}}$, with equality when $d\theta \propto \mathbf{G}^{-1}\nabla_\theta L$. Under the constraint $\|d\theta\|_{\mathbf{G}} = \varepsilon$, the maximal descent direction is $d\theta = -\varepsilon \mathbf{G}^{-1}\nabla_\theta L / \left\|\mathbf{G}^{-1}\nabla_\theta L\right\|_{\mathbf{G}}$. ■

Why natural gradient matters in practice. For overparametrised models (e.g. neural networks, where many different θ give the same predictive distribution), ordinary gradient descent is sensitive to the parametrisation — the step length ε in θ-space does not correspond to a fixed amount of change in the *distribution* p_θ. Natural gradient descent corrects for this by measuring step sizes in the $\mathbf{G}$-metric (which is invariant under reparametrisation of θ, as shown in Section 1.2.4). The practical challenge is computing $\mathbf{G}(\theta)^{-1}$, which costs $O(n_\theta^3)$ for n_θ parameters — infeasible for large neural networks. Kronecker-factored approximate curvature (K-FAC), natural Adam, and similar approximations are active research areas aimed at making natural gradient computationally tractable at scale.

Summary of Section 3.4. The MLE for an exponential family is the e-projection of the empirical distribution $\hat{P}_n$ onto $\mathcal{S}$, characterised by the moment-matching condition $\eta(\hat{\theta}_n) = \hat{\eta}_n$ (Proposition 3.4.1). The Cramér–Rao bound is the Cauchy–Schwarz inequality for the $\mathbf{G}^{-1}$ inner product (Section 3.4.2); the MLE attains it (is first-order efficient) precisely because the e-projection kills the leading-order bias term. Higher-order corrections ($O(1/n)$ bias, Bartlett correction) involve the Levi-Civita curvature built from $T_{ijk}T_{lmn}\mathbf{G}^{-1}\cdots$ contractions (Section 3.4.3). Natural gradient descent is the coordinate-invariant gradient on $(\mathcal{S}, \mathbf{G})$, with the Fisher matrix as the Riemannian metric tensor, and it is invariant under reparametrisation — the geometric content of preconditioning by $\mathbf{G}^{-1}$ (Section 3.4.4).

Curved Exponential Families and the Geometry of Model Misspecification

The preceding three subsections have treated the MLE entirely within a correctly specified exponential family: the true data-generating distribution P^* lies inside the model $\mathcal{S}$, and the MLE converges to the e-projection of $\hat{P}_n$ onto $\mathcal{S}$, which is also the point closest to P^* in KL divergence (by Theorem 3.3.2). This section treats the more realistic scenario of *model misspecification*, in which $P^* \notin \mathcal{S}$, and the conceptually important case of *curved exponential families*, where the model $\mathcal{S}$ is a smooth submanifold of a larger flat exponential family, with positive curvature in the sense of the e-connection.

Curved exponential families. Let $\mathcal{S}_{\text{flat}}$ be a K-dimensional exponential family with natural parameters $\theta \in \Theta \subset \mathbb{R}^K$ and let $\gamma : \Phi \to \Theta$ be a smooth injective map from an open set $\Phi \subset \mathbb{R}^d$ ($d < K$) into the natural parameter space. The *curved exponential subfamily*

$$\mathcal{S}_\gamma = \{p_{\gamma(\varphi)} : \varphi \in \Phi\}$$

is parametrised by φ but embedded in the e-flat ambient manifold $\mathcal{S}_{\text{flat}}$. The induced Fisher metric on $\mathcal{S}_\gamma$ is the pullback $\tilde{g}_{ij}(\varphi) = \frac{\partial \gamma^k}{\partial \varphi^i} \mathbf{G}_{kl}(\gamma(\varphi)) \frac{\partial \gamma^l}{\partial \varphi^j}$, but the induced e-connection on $\mathcal{S}_\gamma$ is *not* flat in general: its curvature tensor is the ambient e-connection curvature (zero, because $\mathcal{S}_{\text{flat}}$ is e-flat) corrected by a *second fundamental form* h^e_{ij} encoding how $\mathcal{S}_\gamma$ curves inside $\mathcal{S}_{\text{flat}}$.

e-Curvature of a Curved Exponential Family

Let $\mathcal{S}_\gamma$ be a d-dimensional curved exponential subfamily of an e-flat K-dimensional family. The *e-curvature* (second fundamental form in the e-connection) at φ is the symmetric bilinear form

$$h^e_{ij}(\varphi) = \Gamma^{(e)\,k}_{ij}(\gamma(\varphi))\, n_k(\varphi),$$

where $\Gamma^{(e)\,k}_{ij}$ are the Christoffel symbols of the e-connection in the ambient family and n_k is the normal component. Equivalently, $h^e_{ij}(\varphi)$ is the component of $\nabla^{(e)}_{\partial_i} \partial_j$ orthogonal to $\mathcal{S}_\gamma$ (in the Fisher-Rao metric). $\mathcal{S}_\gamma$ is e-flat (a linear exponential family) if and only if $h^e_{ij} \equiv 0$.

Canonical example: the Bingham distribution. The simplest non-trivial curved exponential family arises by restricting the mean μ of a multivariate Gaussian to a sphere: the constraint $\|\mu\| = 1$ defines a $(d-1)$-dimensional curved subfamily of the d-dimensional Gaussian location family. Similarly, the Bingham distribution $p(x) \propto \exp(x^\top \mathbf{A} x)$ on the unit sphere S^{d-1} (with $\mathbf{A}$ symmetric, traceless and constrained to have unit Frobenius norm) is a curved exponential family in the larger family of all matrix exponentials on $\mathbb{R}^d$ parametrised by unrestricted symmetric matrices.

Geometry of misspecification. When $P^* \notin \mathcal{S}$, the MLE still converges, but to the point $p_{\hat{\varphi}}$ that minimises $D_{\text{KL}}(P^* \,\|\, p_\varphi)$ over $\varphi \in \Phi$ — the m-projection of P^* onto $\mathcal{S}_\gamma$ (the projection that minimises KL with P^* in the first argument). By Theorem 3.3.1, this is not the same as the e-projection of P^* onto $\mathcal{S}_\gamma$ in general (the e-projection would minimise $D_{\text{KL}}(p_\varphi \,\|\, P^*)$, which is the prediction-risk objective rather than the likelihood objective).

Let $\mathcal{S}_\gamma$ be a curved exponential subfamily and P^* the true distribution. The probability limit of the MLE is

$$\varphi^* = \arg\min_{\varphi \in \Phi} D_{\text{KL}}(P^* \,\|\, p_\varphi) = \pi^{(m)}_{\mathcal{S}_\gamma}(P^*),$$

the m-projection of P^* onto $\mathcal{S}_\gamma$. The MLE is still consistent for this pseudo-true parameter φ^*. The asymptotic variance of $\hat{\varphi}_n$ is

$$\operatorname{Var}(\sqrt{n}(\hat{\varphi}_n - \varphi^*)) \to V(\varphi^*),$$

where V is *not* in general $\tilde{\mathbf{G}}(\varphi^*)^{-1}$ (the Cramér–Rao bound): the sandwich formula $V = \tilde{\mathbf{G}}^{-1} B \tilde{\mathbf{G}}^{-1}$ holds, where $B_{ij} = \operatorname{Cov}_{P^*}[\partial_i \log p_{\varphi^*}, \partial_j \log p_{\varphi^*}]$ is the outer-product Hessian ("bread" of the sandwich) and $\tilde{\mathbf{G}}$ is the model Fisher information. When

$P^* \in \mathcal{S}_\gamma$, $B = \tilde{\mathbf{G}}$ and the sandwich reduces to $\tilde{\mathbf{G}}^{-1}$ (the standard Cramér–Rao formula).

The geometric interpretation is sharp: the m-projection of P^* onto $\mathcal{S}_\gamma$ is the point where the "m-geodesic" from P^* into $\mathcal{S}_\gamma$ is g-orthogonal to $\mathcal{S}_\gamma$. The projection is unique when $\mathcal{S}_\gamma$ is m-convex (i.e. contains all mixture distributions of its elements); for curved exponential families, m-convexity need not hold and the projection may not be unique (the set $\arg\min_\varphi D_{\mathrm{KL}}(P^* \,\|\, p_\varphi)$ may have multiple elements).

Bias from curvature. The leading-order bias $\mathbb{E}[\hat{\varphi}_n] - \varphi^*$ of the MLE in a misspecified or curved exponential family is governed by the e-curvature of $\mathcal{S}_\gamma$ at φ^*. Specifically, Amari (1985, Chapter 4) showed that the $O(1/n)$ bias is proportional to $h^e_{ij}(\varphi^*)\,\tilde{\mathbf{G}}^{ij}(\varphi^*)$, the trace of the e-curvature tensor contracted against the (inverse) Fisher metric. A *flat* model ($h^e \equiv 0$) has zero geometric bias; a curved model has bias proportional to its mean curvature. This is the precise geometric statement behind the heuristic that "parsimonious curved models produce biased MLEs even when the model is correctly specified in a neighbourly sense."

The misspecification setting is not a pathology: most statistical practice involves models that are at best approximate descriptions of reality. The information-geometric framework gives a unified vocabulary for three regimes: (i) $P^* \in \mathcal{S}$ and $\mathcal{S}$ is e-flat (classical MLE, Cramér–Rao efficient, zero curvature); (ii) $P^* \in \mathcal{S}$ and $\mathcal{S}$ is curved (e-curved exponential family, biased MLE, curvature-adjusted asymptotics); (iii) $P^* \notin \mathcal{S}$ (genuine misspecification, sandwich variance, m-projection is the best achievable target). All three regimes are described by the same $(\mathcal{S}, g, \nabla^{(e)}, \nabla^{(m)})$ structure, with curvature and duality as the key geometric invariants.

Numerical Tour: Part 3 — Dual Geodesics, Projections, and Natural Gradient on Bernoulli

This tour makes concrete the abstract geometry of Part 3 for the one-dimensional Bernoulli family, computing e- and m-geodesics between two points, verifying the Pythagorean identity, and comparing ordinary versus natural gradient descent on a simple KL-divergence objective.

Setup. Let $\mathcal{S} = \{\text{Bern}(p) : p \in (0,1)\}$, parametrised by $\theta = \text{logit}(p) = \log(p/(1-p))$ (e-affine coordinate) or $\eta = p$ (m-affine coordinate). Consider two points: P with $p_P = 0.2$ ($\theta_P \approx -1.386$) and Q with $p_Q = 0.7$ ($\theta_Q \approx 0.847$).

Step 1: e-geodesic from P to Q (straight in θ). $\theta(t) = (1-t)\theta_P + t\theta_Q = (1-t)(-1.386) + t(0.847) = -1.386 + 2.233t$. At $t = 0.5$: $\theta(0.5) = -0.2695$, so $p(0.5) = \sigma(-0.2695) = 1/(1+e^{0.2695}) \approx 0.433$. Note: the e-geodesic midpoint at $p \approx 0.433$ is *not* the arithmetic average 0.45 of $p_P = 0.2$ and $p_Q = 0.7$ — it is the *geometric* average of the log-odds.

Step 2: m-geodesic from P to Q (straight in $\eta = p$). $p(t) = (1-t)\cdot 0.2 + t \cdot 0.7 = 0.2 + 0.5t$. At $t = 0.5$: $p(0.5) = 0.45$. The m-geodesic midpoint is exactly the arithmetic average 0.45, as expected from the straight-line structure in η.

Step 3: Pythagorean check. Let $R = \text{Bern}(0.9)$. We check $D_{\text{KL}}(P \,\|\, R) = D_{\text{KL}}(P \,\|\, Q) + D_{\text{KL}}(Q \,\|\, R)$ when the three points are "Pythagorean" (i.e. when Q is the e-projection of P onto an exponential subfamily containing Q and R; for a 1-dimensional family this is trivially never a non-degenerate Pythagorean triple, so we instead verify directly:

$$D_{\text{KL}}(P \,\|\, Q) = 0.2\log\frac{0.2}{0.7} + 0.8\log\frac{0.8}{0.3} = 0.2\times(-1.253) + 0.8\times 0.981 \approx -0.251 + 0.785 = 0.534 \text{ nats.}$$

$$D_{\text{KL}}(Q \,\|\, R) = 0.7\log\frac{0.7}{0.9} + 0.3\log\frac{0.3}{0.1} = 0.7\times(-0.251) + 0.3\times 1.099 \approx -0.176 + 0.330 = 0.154 \text{ nats.}$$

$$D_{\text{KL}}(P \,\|\, R) = 0.2\log\frac{0.2}{0.9} + 0.8\log\frac{0.8}{0.1} = 0.2\times(-1.504) + 0.8\times 2.079 \approx -0.301 + 1.663 = 1.362 \text{ nats.}$$

The triangle inequality holds ($1.362 < 0.534 + 0.154 = 0.688$?), but the Pythagorean identity does *not* hold ($1.362 \neq 0.534 + 0.154$) — as expected: P, Q, R are three arbitrary points, not a Pythagorean triple. The Pythagorean identity holds only when Q is the e- or m-projection of P (or R) onto an appropriate submanifold. In the 1-D case, this submanifold would have to be a single point — a degenerate case. The general Pythagorean theorem becomes non-trivial starting in dimension 2, as in the categorical ($k \geq 3$) and Gaussian (2-dimensional) cases; the numerical examples of Section 2.4.5 (independence constraint) provide the clearest illustration.

Step 4: Natural gradient descent for $\min_p D_{\text{KL}}(\text{Bern}(0.5) \,\|\, \text{Bern}(p))$. Target: $p^* = 0.5$ (minimiser of $L(p) = D_{\text{KL}}(q^* \,\|\, p_\theta)$ with $q^* = \text{Bern}(0.5)$). Starting at $p_0 = 0.1$ ($\theta_0 = \text{logit}(0.1) \approx -2.197$). Ordinary gradient descent in p: $\nabla_p L = -0.5/p + 0.5/(1-p)$; at $p = 0.1$: $-5 + 0.556 = -4.444$. With $\varepsilon = 0.05$: $p_1 = 0.1 + 0.05 \times 4.444 \approx 0.322$. Natural gradient descent in p: $\tilde{\nabla}_p L = \mathbf{G}(p)^{-1}\nabla_p L = p(1-p)\cdot(-0.5/p + 0.5/(1-p)) = 0.5(1-p) - 0.5p = 0.5 - p$; at $p = 0.1$: $0.5 - 0.1 = 0.4$. With $\varepsilon = 0.05$: $p_1 = 0.1 + 0.05 \times 0.4 = 0.12$ (a small, controlled step). In θ-coordinates: $\tilde{\nabla}_\theta L = \mathbf{G}(\theta)^{-1}\nabla_\theta L$; since $\nabla_\theta L = \eta(\theta) - 0.5 = p - 0.5$ (moment-matching condition) and $\mathbf{G}(\theta) = p(1-p)$, $\tilde{\nabla}_\theta L = (p - 0.5)/(p(1-p))$. At $p = 0.1$: $(0.1 - 0.5)/(0.09) \approx -4.44$; update: $\theta_1 = \theta_0 + \varepsilon \times 4.44 = -2.197 + 0.05 \times 4.44 \approx -1.975$, giving $p = \sigma(-1.975) \approx 0.122$. The natural gradient step in θ-space corresponds to exactly the steepest descent in the $\mathbf{G}$-metric on $\mathcal{S}$, and converges in $O(1/\varepsilon)$ steps for this convex problem, whereas ordinary gradient descent in p (which ignores the curvature) overshoots or undershoots depending on p and ε.

Formula Reference: Part 3 — Dual Connections, Flat Geometry, and Natural Gradient

Quantity	Formula / Relation
Amari–Chentsov tensor	$T_{ijk} = \mathbb{E}_\theta[\partial_i \ell\, \partial_j \ell\, \partial_k \ell]$, totally symmetric
T for exp. family	$T_{ijk} = \partial_i \partial_j \partial_k \Phi(\theta)$ (3rd cumulant of sufficient statistic)
α-connection	$\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(0)}_{ij,k} - \frac{\alpha}{2} T_{ijk}$; torsion-free for all α
e-connection ($\alpha = +1$)	$\Gamma^{(e)}_{ij,k} \equiv 0$ in θ-coordinates for any exp. family
m-connection ($\alpha = -1$)	$\tilde{\Gamma}^{(m)}_{ij,k} \equiv 0$ in η-coordinates for any exp. family
Duality	$\nabla^{(\alpha)}, \nabla^{(-\alpha)}$ are dual w.r.t. $\mathbf{G}$: $\partial_k g_{ij} = \Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i}$
e-geodesics	Lines in θ: $\theta(t) = (1-t)\theta_0 + t\theta_1$ (exponential arcs: $p_{\theta(t)} \propto p_0^{1-t} p_1^t$)
m-geodesics	Lines in η: $\eta(t) = (1-t)\eta_0 + t\eta_1$ (arithmetic mixtures: $p_t = (1-t)p_0 + t p_1$)
Dual potentials	$\Phi(\theta) + \Psi(\eta) = \theta^\top \eta$ (Young–Fenchel); $\eta = \nabla\Phi$, $\theta = \nabla\Psi$
Metric from potentials	$g_{ij} = \partial_i \partial_j \Phi(\theta)$; $g^{ij} = \partial^i \partial^j \Psi(\eta)$
Canonical divergence	$D(P\|Q) = \Phi(\theta_P) + \Psi(\eta_Q) - \theta_P^\top \eta_Q = B_\Phi(\theta_P : \theta_Q) = D_{\mathrm{KL}}(p_P \,\|\, p_Q)$
Geometric Pythagorean	$D(P\|R) = D(P\|Q) + D(Q\|R)$ iff m-geodesic PQ $\perp_{\mathbf{G}}$ e-geodesic QR at Q
MLE as e-projection	$\hat{\theta}_n = \arg\min_\theta D_{\mathrm{KL}}\big(\hat{P}_n \,\|\, p_\theta\big)$; characterised by $\eta(\hat{\theta}_n) = \hat{\eta}_n$
Moment-matching	$\mathbb{E}_{\hat{\theta}_n}[T(X)] = \frac{1}{n}\sum_i T(x_i)$: MLE = method of moments for exp. families
Cramér–Rao bound	$\mathrm{Cov}_\theta(\hat{\theta}) \succeq \mathbf{G}(\theta)^{-1}$ for unbiased $\hat{\theta}$; MLE is efficient (attains bound)
L-C curvature (exp. fam.)	$R^{(0)}_{kl,ij} = -\frac{1}{4}(T_{kia} g^{ab} T_{ljb} - T_{kja} g^{ab} T_{lib})$ (nonzero unless $T = 0$)
e/m flatness	$R^{(e)} = R^{(m)} = 0$ for every exponential family; $R^{(0)} = 0$ iff $T \equiv 0$
Natural gradient	$\tilde{\nabla}_\theta L = \mathbf{G}(\theta)^{-1} \nabla_\theta L$ (steepest descent in $\mathbf{G}$-metric)
Natural gradient descent	$\theta_{t+1} = \theta_t - \varepsilon \mathbf{G}(\theta_t)^{-1} \nabla_\theta L(\theta_t)$; parametrisation-invariant
Bartlett correction	$c = 1 + b/n$; b depends on $T_{ijk}\mathbf{G}^{-1}T_{lmn}$-contractions; $n\Lambda_n/c \to \chi^2_k$ more accurately

Part 4 — Information Projections and the EM Algorithm

The e-projection and m-projection as minimisers of KL divergence. The generalised Pythagorean theorem as the foundation of alternating projections. Csiszár–Tusnády analysis: convergence of alternating e/m-projections. The EM algorithm as alternating projections on a dually flat manifold. Iterative proportional fitting and its information-geometric interpretation. Applications: Gaussian mixture models, hidden Markov models, and exponential family EM.

4.1 Information Projections: e- and m-Projections

The Two Fundamental Projections

Let $\mathcal{S} = \{p_\theta : \theta \in \Theta\}$ be an exponential family with natural coordinates θ and moment coordinates η, and let $\mathcal{E} \subseteq \mathcal{S}$ be a submanifold. Two natural notions of "closest point in $\mathcal{E}$" arise from the two flat structures.

Definition 4.1 — e- and m-Projections

Let q be a distribution and $\mathcal{E}$ a submanifold of a dually flat space.

(i) The *e-projection* (exponential projection) of q onto $\mathcal{E}$ is

$$\pi^{(e)}(q \to \mathcal{E}) = \arg\min_{p \in \mathcal{E}} D_{\mathrm{KL}}(q \,\|\, p)\,.$$

(ii) The *m-projection* (mixture projection) of q onto $\mathcal{E}$ is

$$\pi^{(m)}(q \to \mathcal{E}) = \arg\min_{p \in \mathcal{E}} D_{\mathrm{KL}}(p \,\|\, q)\,.$$

The asymmetry of KL divergence is reflected in these two different projections. The e-projection minimises $D_{\mathrm{KL}}(q \,\|\, \cdot)$, penalising $p \in \mathcal{E}$ heavily if $p(x)$ is small where $q(x)$ is large ("mode-seeking" or "zero-forcing" behaviour). The m-projection minimises $D_{\mathrm{KL}}(\cdot \,\|\, q)$, which has a different behaviour: it penalises p if $p(x)$ is large where $q(x)$ is small ("mean-seeking" or "inclusive" behaviour).

Theorem 4.1 — Geometric Characterisation of Projections

In a dually flat space with potential ψ (in θ-coordinates):

(i) If $\mathcal{E}$ is an e-flat submanifold (flat in θ-coordinates), the m-projection $\pi^{(m)}(q \to \mathcal{E})$ exists and is unique; it is the point $p^* \in \mathcal{E}$ where the e-geodesic from p^* to q meets $\mathcal{E}$ orthogonally (in the Fisher metric).

(ii) If $\mathcal{E}$ is an m-flat submanifold (flat in η-coordinates), the e-projection $\pi^{(e)}(q \to \mathcal{E})$ exists and is unique; it is the point $p^* \in \mathcal{E}$ where the m-geodesic from p^* to q meets $\mathcal{E}$ orthogonally.

Proof of (ii). The e-projection solves $\min_{p \in \mathcal{E}} D_{\mathrm{KL}}(q \,\|\, p)$. Write the KL divergence using the Bregman representation on the dually flat space:

$$D_{\mathrm{KL}}(q \,\|\, p) = \psi(\theta_p) + \psi^*(\eta_q) - \langle \theta_p, \eta_q \rangle = B_\psi(\theta_q \| \theta_p) + \text{const.}$$

This is a Bregman divergence in θ_p, which is strictly convex. If $\mathcal{E}$ is m-flat, it is an affine subspace in η-coordinates, which corresponds (via the Legendre map $\nabla\psi$) to a convex set in θ-coordinates. A Bregman divergence minimised over a convex set has a unique minimum,

characterised by the first-order condition: $\langle \nabla_\theta B_\psi(\theta_q \| \theta_{p^*}), \theta_p - \theta_{p^*} \rangle \geq 0$ for all $p \in \mathcal{E}$. This is precisely the orthogonality condition in the Fisher metric. ■

Pythagorean Decomposition and Projection Error

The Pythagorean theorem for Bregman divergences (proved in Part 2) becomes, for projections, a decomposition of the total divergence from a reference point:

Theorem 4.2 — Pythagorean Decomposition of Projection

Let r be any distribution, $\mathcal{E}$ be an m-flat submanifold, and $p^* = \pi^{(e)}(r \to \mathcal{E})$. Then for every $p \in \mathcal{E}$:

$$D_{\mathrm{KL}}(r \,\|\, p) = D_{\mathrm{KL}}(r \,\|\, p^*) + D_{\mathrm{KL}}(p^* \,\|\, p)\,.$$

Equivalently: $D(r\|p) = D(r\|p^*) + D(p^*\|p)$, with the first term being the minimum achievable divergence (the "projection error") and the second being the "extra" divergence due to not being at the optimal point.

This identity is the engine of the EM algorithm: it shows that improving $D_{\mathrm{KL}}(r \,\|\, p)$ by moving p within $\mathcal{E}$ reduces $D_{\mathrm{KL}}(p^* \,\|\, p)$ without touching the irreducible term $D_{\mathrm{KL}}(r \,\|\, p^*)$.

Example — e-Projection onto a Gaussian Family

Let r be an arbitrary distribution on $\mathbb{R}$ with mean μ_r and variance σ_r^2. The Gaussian family $\mathcal{E} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ is an exponential family with natural parameters $\theta = (\theta^1, \theta^2) = (\mu/\sigma^2, -1/(2\sigma^2))$ and sufficient statistics $T(x) = (x, x^2)$. Since $\mathcal{E}$ is exponential (hence e-flat, but also m-flat in its own right as a 2-dimensional family with e-flat submanifolds), the e-projection of r onto $\mathcal{E}$ is the Gaussian $p^* = \mathcal{N}(\mu_r, \sigma_r^2)$ that matches the mean and variance of r.
This follows from the moment-matching characterisation: the e-projection satisfies $\mathbb{E}_{p^*}[T(X)] = \mathbb{E}_r[T(X)]$, i.e., $(\mu_{p^*}, \mathbb{E}_{p^*}[X^2]) = (\mu_r, \mu_r^2 + \sigma_r^2)$. The KL divergence $D_{\mathrm{KL}}(r \,\|\, \mathcal{N}(\mu_r, \sigma_r^2))$ equals the negative differential entropy of r minus $-H(\mathcal{N}(\mu_r, \sigma_r^2))$, and it is minimised precisely at $p^* = \mathcal{N}(\mu_r, \sigma_r^2)$.

4.2 The EM Algorithm as Alternating Projections

The EM Algorithm: Setup and Information-Geometric Reformulation

The *Expectation-Maximisation* algorithm, introduced by Dempster, Laird, and Rubin (1977), is a widely used procedure for computing maximum likelihood estimates in the presence of latent (unobserved) variables. Its information-geometric structure — as alternating projections on a dually flat manifold — was first elucidated by Amari (1995) and provides a unified analysis of its convergence.

Let X be the observed data and Z the latent variable, with complete-data distribution $p_\theta(x, z) = p_\theta(z|x)p_\theta(x)$ in an exponential family parametrised by θ. The observed-data log-likelihood $\ell(\theta) = \log p_\theta(x) = \log \int p_\theta(x, z)\, dz$ is generally intractable. The EM algorithm introduces an auxiliary distribution $q(z)$ over the latent variable and iterates:

Step 1: **E-step**: Given current parameter $\theta^{(t)}$, compute the conditional distribution $q^{(t+1)}(z) = p_{\theta^{(t)}}(z|x)$.

Step 2: **M-step**: Update $\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{q^{(t+1)}}[\log p_\theta(X, Z)]$.

Information-geometric reformulation. Consider the manifold $\mathcal{M} = \{p_\theta(x,z) : \theta \in \Theta\}$ (the model for complete data) and the submanifold $\mathcal{F} = \{q(x,z) : q(z|x) = p_{\hat{\theta}}(z|x) \text{ for some } \hat{\theta}\}$ (distributions that are "consistent with the observed data" in that the x-marginal is an empirical observation). The EM algorithm performs:

- **E-step** = m-projection of the current model onto the data-consistent submanifold $\mathcal{F}$.
- **M-step** = e-projection of the data-consistent distribution back onto the model $\mathcal{M}$.

Each step reduces the KL divergence to the empirical distribution, and the Pythagorean theorem guarantees that the total divergence decreases monotonically.

Convergence Analysis via the Csiszár–Tusnády Theorem

Theorem 4.3 — Csiszár–Tusnády / EM Monotonicity

Let $\mathcal{M}$ and $\mathcal{F}$ be e-flat and m-flat submanifolds respectively of a dually flat space, intersecting at a unique point p^*. Starting from any $p^{(0)} \in \mathcal{M}$, the alternating projection sequence

$$q^{(t+1)} = \pi^{(m)}(p^{(t)} \to \mathcal{F}), \qquad p^{(t+1)} = \pi^{(e)}(q^{(t+1)} \to \mathcal{M})$$

satisfies:

$$D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t+1)}\big) \le D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t)}\big) \quad \text{for all } t,$$

with equality only at convergence $p^{(t)} = p^*$. Moreover, the sequence converges to p^* at a linear rate determined by the second-order geometry of the intersection.

Proof sketch. By the Pythagorean decomposition at $q^{(t+1)}$:

$$D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t)}\big) = D_{\mathrm{KL}}\big(p^* \,\|\, q^{(t+1)}\big) + D_{\mathrm{KL}}\big(q^{(t+1)} \,\|\, p^{(t)}\big)\,.$$

Since $p^* \in \mathcal{F}$ and $q^{(t+1)} = \pi^{(m)}(p^{(t)} \to \mathcal{F})$, the Pythagorean theorem for m-projection gives $D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t)}\big) \ge D_{\mathrm{KL}}\big(p^* \,\|\, q^{(t+1)}\big)$. Applying the same argument to the e-projection step:

$$D_{\mathrm{KL}}\big(p^* \,\|\, q^{(t+1)}\big) \ge D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t+1)}\big)\,.$$

Chaining these gives $D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t+1)}\big) \le D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t)}\big)$. Strict inequality holds unless $p^{(t)} = p^*$, since the Pythagorean remainder $D_{\mathrm{KL}}\big(q^{(t+1)} \,\|\, p^{(t)}\big)$ is zero if and only if $p^{(t)} \in \mathcal{F}$, which happens only at the intersection. ■

Exponential Family EM: Explicit Formulae

When $p_\theta(x,z) = \exp(\langle \theta, T(x,z)\rangle - \psi(\theta))$ is an exponential family in the complete data (x,z), the M-step simplifies dramatically. Let $\eta = \nabla\psi(\theta)$ be the moment parameters and T the sufficient statistics.

E-step computes the conditional expected sufficient statistics:

$$\bar{T}^{(t+1)} = \mathbb{E}_{q^{(t+1)}}[T(X,Z)] = \mathbb{E}_{p_{\theta^{(t)}}(\cdot|x)}[T(x,Z)].$$

M-step finds $\theta^{(t+1)}$ such that the moment parameters match:

$$\eta(\theta^{(t+1)}) = \bar{T}^{(t+1)}, \quad \text{i.e.,} \quad \theta^{(t+1)} = (\nabla\psi)^{-1}(\bar{T}^{(t+1)}) = \nabla\psi^*(\bar{T}^{(t+1)}).$$

This is the moment-matching equation: the MLE in an exponential family sets the expected sufficient statistics equal to their observed value. When Z is latent, we use the conditional expectation under the current model — this is precisely the EM iteration.

Example — Gaussian Mixture Model via EM

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2)$, where the component assignments $Z_i \in \{1, \dots, K\}$ are latent. Parameter: $\theta = (\pi, \mu, \sigma^2)$.
E-step: Compute the *responsibilities*

$$\gamma_{ik}^{(t)} = \frac{\pi_k^{(t)} \, \phi(x_i; \mu_k^{(t)}, \sigma_k^{2(t)})}{\sum_{l=1}^{K} \pi_l^{(t)} \, \phi(x_i; \mu_l^{(t)}, \sigma_l^{2(t)})},$$

where $\phi(\cdot; \mu, \sigma^2)$ is the Gaussian density. These are the posterior probabilities $p(Z_i = k | X_i = x_i, \theta^{(t)})$.
M-step: Update parameters using weighted moments:

$$\begin{aligned}
N_k^{(t+1)} &= \sum_{i=1}^{n} \gamma_{ik}^{(t)}, \\
\pi_k^{(t+1)} &= \frac{N_k^{(t+1)}}{n}, \\
\mu_k^{(t+1)} &= \frac{1}{N_k^{(t+1)}} \sum_{i=1}^{n} \gamma_{ik}^{(t)} \, x_i, \\
\sigma_k^{2(t+1)} &= \frac{1}{N_k^{(t+1)}} \sum_{i=1}^{n} \gamma_{ik}^{(t)} \, (x_i - \mu_k^{(t+1)})^2.
\end{aligned}$$

The information-geometric interpretation: $\gamma_{ik}^{(t)}$ defines the m-projection of the empirical distribution onto the mixture manifold at $\theta^{(t)}$ (E-step), and the M-step performs the e-projection back onto the exponential family component of the complete-data model. Convergence to a local maximum of the observed likelihood is guaranteed by the Csiszár–Tusnády theorem.

Convergence Rate of EM and the Local Geometry at the Fixed Point

The Csiszár–Tusnády theorem guarantees monotone decrease of the KL divergence, but does not quantify the convergence rate. The rate is governed by the geometry of the intersection $p^* \in \mathcal{M} \cap \mathcal{F}$ and, for the standard EM algorithm, is characterised by the *fraction of missing information*.

Theorem 4.4 — EM Rate via Missing Information

For an exponential family EM with complete-data Fisher information $\mathbf{G}_c(\theta^*)$ and observed-data Fisher information $\mathbf{G}_o(\theta^*)$, the convergence rate of EM at the fixed point θ^* is determined by the *matrix rate*

$$R_{\text{EM}} = I - \mathbf{G}_o(\theta^*)\mathbf{G}_c(\theta^*)^{-1} = \mathbf{G}_c(\theta^*)^{-1}\mathbf{G}_m(\theta^*),$$

where $\mathbf{G}_m(\theta^*) = \mathbf{G}_c(\theta^*) - \mathbf{G}_o(\theta^*)$ is the *missing information* (the Fisher information lost by not observing Z). Specifically:

$$\|\theta^{(t+1)} - \theta^*\| \leq \|R_{\text{EM}}\| \cdot \|\theta^{(t)} - \theta^*\| + O(\|\theta^{(t)} - \theta^*\|^2),$$

so the asymptotic linear convergence rate is $\lambda_{\max}(R_{\text{EM}}) = 1 - \lambda_{\min}(\mathbf{G}_o\mathbf{G}_c^{-1})$.

The formula $R_{\text{EM}} = \mathbf{G}_c^{-1}\mathbf{G}_m$ has a transparent information-geometric interpretation: $\mathbf{G}_c^{-1}$ is the complete-data metric (how fast the geometry moves per unit change in θ), while $\mathbf{G}_m$ is the contribution of the *missing* data to the curvature. When Z is fully observed ($\mathbf{G}_m = 0$), EM becomes one-step Newton and converges quadratically. When the latent variable Z carries most of the information about θ ($\mathbf{G}_m \approx \mathbf{G}_c$, i.e. $\mathbf{G}_o \approx 0$), $R_{\text{EM}} \approx I$ and EM converges arbitrarily slowly.

In the submanifold picture: R_{EM} is (proportional to) $\cos^2 \angle(\mathcal{M}, \mathcal{F})$ at p^*, the squared cosine of the angle between the model manifold $\mathcal{M}$ and the data-consistent manifold $\mathcal{F}$ in the Fisher metric at the solution. This is exactly the same formula as the IPF rate $c^2 = \lambda_2(M)^2$ of Section 4.3, making explicit that EM and IPF are the same geometric operation (alternating e/m-projections) applied to different pairs of submanifolds.

Example — EM Rate for a Two-Component Gaussian Mixture

For $X \sim \pi_1\mathcal{N}(\mu_1, 1) + (1-\pi_1)\mathcal{N}(\mu_2, 1)$ with $\pi_1 = 0.5$ fixed (so $\theta = (\mu_1, \mu_2)$), the EM rate matrix at a symmetric fixed point $\mu_1^* = -\mu_2^* = \delta > 0$ is

$$R_{\text{EM}} = \begin{pmatrix} 1-r & -r \\ -r & 1-r \end{pmatrix}, \quad r = \frac{1}{4}\mathbb{E}_{p^*}\Big[\operatorname{sech}^2\Big(\delta \cdot (X - \delta)\Big)\Big],$$

where the expectation is over $X \sim p^*$. The scalar rate is $\lambda_{\max}(R_{\text{EM}}) = 1 - 2r$. For small separation $\delta \to 0$: $r \approx 1/4$ and rate $\approx 1/2$ (slow, 50% reduction per step). For large separation $\delta \to \infty$: $r \approx 0$ and rate ≈ 1 (very slow, near-zero reduction per step) because the two components become perfectly distinguishable ($\mathbf{G}_o \approx \mathbf{G}_c$, $\mathbf{G}_m \approx 0$) — but wait, this seems backwards. In fact for large δ, $\operatorname{sech}^2(\delta(x-\delta)) \to 0$ for all x near either component, so $r \to 0$ and the rate $\to 1$: this is the case where EM is near its fixed point with well-separated components, but the *convergence is still linear* because the iterates oscillate between the two symmetric fixed points at $\pm\delta$ if not initialised correctly. The slow convergence for large δ reflects the flat geometry of the log-likelihood near the (local) MLE, not fast convergence.

Generalised EM, ECM, and ECME

The geometric framework immediately suggests generalisations of EM that may converge faster or handle more complex latent structures.

Generalised EM (GEM). The M-step need not find the exact maximiser of the Q-function; it suffices to find any $\theta^{(t+1)}$ satisfying $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$ (an *ascent* in the Q-function, not a maximum). The monotonicity property of Theorem 4.3 is preserved: $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$. In the geometric picture, GEM replaces the exact e-projection step with any step that reduces the KL divergence to $\mathcal{F}$, and the Pythagorean monotonicity still holds.

ECM (Expectation Conditional Maximisation, Meng and Rubin 1993). When the M-step is computationally expensive, ECM replaces it with a sequence of *conditional* M-steps, each maximising the Q-function over one block of parameters with the others fixed:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}) \text{ s.t. } g_j(\theta) = c_j^{(t)}, j \neq \text{current}.$$

Geometrically, ECM is an alternating e-projection that visits a sequence of coordinate submanifolds, converging to the same limit as EM but with each step being computationally cheaper.

ECME (Liu and Rubin 1994). ECME modifies ECM to allow some conditional M-steps to

maximise the *observed* likelihood $\ell(\theta)$ rather than the Q-function $Q(\theta|\theta^{(t)})$. This can dramatically accelerate convergence for parameters that appear in both the complete-data likelihood and the observed-data likelihood in a simple way (e.g., a mixing proportion or a dispersion parameter). The trade-off is a more complex implementation of the M-step.

The information-geometric analysis of EM (and its variants) is particularly revealing for the *hidden Markov model* (HMM). In an HMM with hidden state sequence $Z = (Z_1, \dots, Z_T)$ and observations $X = (X_1, \dots, X_T)$, the complete-data sufficient statistics are the transition counts $\{n_{ij}\}$ and the emission statistics $\{s_k(x_t)\}$. The EM algorithm for HMMs (Baum-Welch algorithm) computes the E-step via the forward-backward algorithm (computing $\mathbb{E}[n_{ij}|X, \theta^{(t)}]$ and $\mathbb{E}[s_k(X_t)|X, \theta^{(t)}]$ in $O(T \cdot K^2)$ operations, where K is the number of states) and the M-step via closed-form moment-matching. The convergence rate is governed by the spectral gap of the transition matrix: slow mixing of the hidden Markov chain corresponds to high missing information ($\mathbf{G}_m$ large relative to $\mathbf{G}_c$), and hence slow EM convergence. This connects the information-geometric rate formula to the spectral theory of Markov chains, unifying two apparently different theories under the dually-flat projection framework.

EM algorithm summary. Q-function: $Q(\theta|\theta') = \mathbb{E}_{p_{\theta'}(Z|X=x)}[\log p_\theta(X, Z)]$. E-step ($m$-projection): $q^{(t+1)} = p_{\theta^{(t)}}(\cdot|x)$. M-step ($e$-projection): $\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)})$. Monotonicity: $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$ (by Pythagorean theorem). Rate: $\lambda_{\max}(I - \mathbf{G}_o\mathbf{G}_c^{-1})$ (missing-information fraction). Exp. family M-step: $\eta(\theta^{(t+1)}) = \mathbb{E}_{p_{\theta^{(t)}}(\cdot|x)}[T(X, Z)]$ (moment matching).

4.3 Iterative Proportional Fitting and Contingency Tables

The IPF Algorithm

Iterative proportional fitting (IPF), also known as the Sinkhorn-Knopp algorithm in the matrix case, solves the problem of finding a distribution with prescribed marginals that is closest (in KL divergence) to a given reference distribution. This is a canonical example of alternating m-projections.

Definition 4.2 — Marginal Projection Problem

Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and let q be a joint distribution. Given target marginals $r_1 \in \mathcal{P}(\mathcal{X}_1)$ and $r_2 \in \mathcal{P}(\mathcal{X}_2)$, the IPF problem is:

$$\min_{p \in \mathcal{C}} D_{\mathrm{KL}}(p \,\|\, q), \qquad \mathcal{C} = \{p : p_{X_1} = r_1,\ p_{X_2} = r_2\}.$$

The constraint set $\mathcal{C}$ is the intersection of two m-flat submanifolds: $\mathcal{C}_1 = \{p : p_{X_1} = r_1\}$ and $\mathcal{C}_2 = \{p : p_{X_2} = r_2\}$. IPF alternates m-projections:

$$p^{(t+1/2)} = \pi^{(m)}(p^{(t)} \to \mathcal{C}_1): \quad p^{(t+1/2)}(x_1, x_2) = p^{(t)}(x_1, x_2) \cdot \frac{r_1(x_1)}{p^{(t)}_{X_1}(x_1)},$$

$$p^{(t+1)} = \pi^{(m)}(p^{(t+1/2)} \to \mathcal{C}_2): \quad p^{(t+1)}(x_1, x_2) = p^{(t+1/2)}(x_1, x_2) \cdot \frac{r_2(x_2)}{p^{(t+1/2)}_{X_2}(x_2)}.$$

Each half-step is a simple ratio rescaling. Convergence follows from the Bregman projection

theorem: alternating m-projections onto closed convex sets in KL geometry converge to the point in $\mathcal{C}_1 \cap \mathcal{C}_2$ that minimises $D_{\mathrm{KL}}(p \,\|\, q)$ (Csiszár 1975).

Solution structure. The unique solution to the IPF problem has the form $p^*(x_1, x_2) = q(x_1, x_2) \cdot u(x_1) \cdot v(x_2)$ for some scaling functions $u : \mathcal{X}_1 \to \mathbb{R}_{>0}$ and $v : \mathcal{X}_2 \to \mathbb{R}_{>0}$. This is the *I-projection of q with prescribed marginals*, and it is the maximum-entropy distribution consistent with the marginals and the independence structure of q.

In optimal transport, IPF is the algorithmic workhorse for computing regularised optimal transport plans: given a cost matrix C and marginals r_1, r_2, the Sinkhorn algorithm with entropic regularisation $\epsilon > 0$ solves $\min_p \langle C, p\rangle + \epsilon D_{\mathrm{KL}}(p \,\|\, \mathbf{1})$ s.t. $p_{X_1} = r_1$, $p_{X_2} = r_2$, and is precisely IPF with reference $q(x_1, x_2) = \exp(-C(x_1, x_2)/\epsilon)/Z$.

Convergence Rate of IPF

The convergence rate of alternating m-projections is determined by the angle between the two submanifolds $\mathcal{C}_1$ and $\mathcal{C}_2$ in the Fisher metric at the solution p^*. Define the *cosine of the angle* as

$$c = \sup \left\{ \frac{g_{p^*}(u, v)}{\|u\|_{p^*} \|v\|_{p^*}} : u \in T_{p^*}\mathcal{C}_1,\ v \in T_{p^*}\mathcal{C}_2,\ u, v \neq 0 \right\}.$$

Then IPF converges linearly with rate c^2:

$$D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t+1)}\big) \leq c^2 \cdot D_{\mathrm{KL}}\big(p^* \,\|\, p^{(t)}\big)\,.$$

When $\mathcal{C}_1 \perp \mathcal{C}_2$ at p^* (i.e., $c = 0$, which happens when the marginals are independent under p^*), IPF converges in *one step*: the product distribution $r_1 \otimes r_2$ is already the minimiser. In general, the submanifolds are not orthogonal and the rate $c^2 < 1$ governs the linear convergence.

Spectral interpretation. For a $K \times K$ contingency table problem with marginal constraints $\mathcal{C}_1 = \{p : p_X = r_X\}$ and $\mathcal{C}_2 = \{p : p_Y = r_Y\}$, the convergence rate c^2 of IPF equals the square of the second-largest singular value of the matrix M with entries $M_{xy} = p^*(x, y)/\sqrt{p^*(x, y) r_X(x) r_Y(y)}$ (the normalised contingency matrix). Specifically, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K$ be the singular values of M; then $\lambda_1 = 1$ always (the first singular vector corresponds to the projection direction onto $C_1 \cap C_2$) and

$$c = \lambda_2(M), \qquad \text{convergence rate} = \lambda_2(M)^2.$$

When $p^* = r_X \otimes r_Y$ (the independent product), M is the all-ones matrix with rank 1, so $\lambda_2 = 0$ and convergence is one-step. When p^* has strong dependence (near-deterministic association between X and Y), $\lambda_2 \approx 1$ and IPF converges slowly.

Example — Convergence Rate for a 2×2 Table

Let $p^* = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with $a + b = r_X(1)$, $c + d = r_X(2)$, $a + c = r_Y(1)$, $b + d = r_Y(2)$. The normalised matrix $M_{xy} = p^*(x, y)/\sqrt{p^*(x, y) r_X(x) r_Y(y)} = \sqrt{p^*(x, y)/(r_X(x) r_Y(y))}$. The second singular value of a rank-2 matrix M with constant row/column sums (since M has both row sums and column sums equal to 1 by the marginal constraint) is $\lambda_2 = |\mathrm{corr}_{p^*}(\phi(X), \psi(Y))|$ for the optimal pair of functions ϕ, ψ (the first canonical correlation coefficient between X and Y under p^*). For the case $a = d = 0.4$, $b = c = 0.1$ (positively correlated table with $r_X = r_Y = (0.5, 0.5)$): $M = \left(\begin{smallmatrix} \sqrt{0.4/0.25} & \sqrt{0.1/0.25} \\ \sqrt{0.1/0.25} & \sqrt{0.4/0.25} \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1.265 & 0.632 \\ 0.632 & 1.265 \end{smallmatrix}\right)$, singular values $\lambda_1 = 1.897/\sqrt{2} = 1$, $\lambda_2 = 0.633/\sqrt{2} \approx 0.447$. IPF convergence rate $c^2 \approx 0.20$: each iteration reduces the KL gap by about 80%.

The quantity $c = \lambda_2(M)$ has two equivalent geometric interpretations:

1. **Submanifold angle.** $c = \cos \angle(T_{p^*}\mathcal{C}_1, T_{p^*}\mathcal{C}_2)$ in the Fisher metric g_{p^*} (the cosine of the principal angle between the tangent spaces to the two constraint submanifolds at the solution).
2. **Maximal canonical correlation.** $c = \max_{\phi,\psi} \mathrm{Corr}_{p^*}(\phi(X), \psi(Y))$, the maximal canonical correlation (Hirschfeld 1935, Gebelein 1941) between X and Y under p^*. This is also called the *Hirschfeld–Gebelein–Rényi maximal correlation* $M(X;Y)$ and satisfies $0 \le M(X;Y) \le 1$, with $M(X;Y) = 0$ iff $X \perp Y$ under p^* and $M(X;Y) = 1$ iff X and Y are functionally dependent.

Thus IPF converges in one step iff X and Y are p^*-independent, and converges arbitrarily slowly (rate $\to 1$) iff X and Y are p^*-nearly-determined by each other.

Multi-Marginal IPF and Entropy-Regularised Optimal Transport

The two-marginal IPF algorithm extends naturally to the *multi-marginal* setting, where one seeks a joint distribution p over $\mathcal{X}_1 \times \cdots \times \mathcal{X}_K$ with prescribed marginals $r_k \in \mathcal{P}(\mathcal{X}_k)$, $k = 1, \dots, K$, closest to a reference q in KL divergence:

$$\min_{p \in \mathcal{C}} D_{\mathrm{KL}}(p \,\|\, q), \qquad \mathcal{C} = \bigcap_{k=1}^{K} \mathcal{C}_k, \quad \mathcal{C}_k = \{p : p_{X_k} = r_k\}.$$

Each $\mathcal{C}_k$ is m-flat (a linear family in the expectation parameters of the marginal functions $\mathbf{1}_{X_k = x_k}$). The alternating m-projection algorithm cycles through $k = 1, \dots, K$ and at each step rescales p so that its k-th marginal matches r_k:

$$p^{(t,k)}(x) = p^{(t,k-1)}(x) \cdot \frac{r_k(x_k)}{p^{(t,k-1)}_{X_k}(x_k)},$$

with the convention $p^{(t,0)} = p^{(t-1,K)}$. The solution always has the *Gibbs form*

$$p^*(x_1, \dots, x_K) = q(x_1, \dots, x_K) \cdot \prod_{k=1}^{K} u_k(x_k)$$

for scaling potentials $u_k : \mathcal{X}_k \to \mathbb{R}_{>0}$, a direct generalisation of the two-marginal case.

Entropy-regularised optimal transport. The deepest application of IPF in modern computational mathematics is the Sinkhorn algorithm for entropy-regularised optimal transport. Given two discrete distributions $r_1 \in \mathcal{P}(\mathcal{X}_1)$ and $r_2 \in \mathcal{P}(\mathcal{X}_2)$ and a cost matrix $C \in \mathbb{R}^{|\mathcal{X}_1| \times |\mathcal{X}_2|}$ (where $C(x_1, x_2)$ is the cost of transporting a unit mass from x_1 to x_2), the Kantorovich optimal transport problem is:

$$\min_{p \in \Pi(r_1, r_2)} \langle C, p \rangle \;=\; \min_{p \in \Pi(r_1, r_2)} \sum_{x_1, x_2} C(x_1, x_2)\, p(x_1, x_2),$$

where $\Pi(r_1, r_2) = \{p : p_{X_1} = r_1, p_{X_2} = r_2\}$ is the set of *transport plans* (couplings with the right marginals). This is a linear program, but its computational cost is $O(n^3)$ (by the network simplex algorithm) for distributions supported on n points — too expensive for large-scale applications (e.g. computing the Wasserstein distance between two point clouds of size 10^4).

The entropy-regularised relaxation adds a KL penalty:

$$\min_{p \in \Pi(r_1, r_2)} \langle C, p \rangle + \varepsilon D_{\mathrm{KL}}(p \,\|\, r_1 \otimes r_2)\,.$$

The unique solution satisfies the first-order condition $p^*(x_1, x_2) \propto \exp(-C(x_1, x_2)/\varepsilon) \cdot u(x_1) \cdot v(x_2)$, which is precisely the Gibbs form of the IPF solution with reference distribution $q(x_1, x_2) = \exp(-C(x_1, x_2)/\varepsilon)/(Z_\varepsilon)$ (unnormalised). The Sinkhorn algorithm is therefore IPF applied to this reference, and the information-geometric reading is immediate: the regularised transport plan is the m-projection of $q \propto e^{-C/\varepsilon}$ onto the marginal constraints $\Pi(r_1, r_2)$.

The Sinkhorn algorithm with regularisation $\varepsilon > 0$ is the alternating m-projection algorithm applied to the reference $q = e^{-C/\varepsilon}/Z$ and the constraint sets $\mathcal{C}_1 = \{p : p_{X_1} = r_1\}$, $\mathcal{C}_2 = \{p : p_{X_2} = r_2\}$. It converges linearly with rate c^2 (the cosine-squared of the angle between $\mathcal{C}_1$ and $\mathcal{C}_2$ at p^*), and the rate c^2 improves (decreases) as $\varepsilon \to 0$ (because the optimal transport plan concentrates mass near the optimal matching, making the two marginal constraints nearly orthogonal at the solution).

Remark on the limit $\varepsilon \to 0$. As $\varepsilon \to 0$, the entropy-regularised solution p^*_ε concentrates on the support of an optimal (unregularised) transport plan, and the Sinkhorn iterates converge to a discrete approximation of the classical Kantorovich solution. The information-geometric perspective is that the optimal transport problem itself — as a limit — corresponds to finding the m-projection of the degenerate measure $e^{-C/\varepsilon}/Z$ (which, as $\varepsilon \to 0$, concentrates on the cost-minimising support) onto the marginal constraint set. The geometry of the convergence, including the rate of divergence of the number of Sinkhorn iterations as $\varepsilon \to 0$, is governed by the curvature of the constraint manifolds at their intersection point.

Multi-marginal Sinkhorn and Wasserstein barycentres. The multi-marginal generalisation of Sinkhorn computes *Wasserstein barycentres*: given K distributions $r_1, \dots, r_K$ and positive weights $\lambda_1, \dots, \lambda_K$ (with $\sum \lambda_k = 1$), the Wasserstein barycentre is the distribution μ^* that minimises $\sum_{k=1}^K \lambda_k W_2^2(\mu, r_k)$ (the weighted sum of squared Wasserstein distances). Cuturi and Doucet (2014) showed that this problem can be reformulated as a multi-marginal optimal transport problem with reference $q(x_1, \dots, x_K) = \exp(-\sum_k \lambda_k C(x_k, \cdot)/\varepsilon)$ and a cost that encodes the common "barycentre location"; the multi-marginal Sinkhorn algorithm then computes the solution by alternating projections onto each of the K marginal constraint sets.

Computational cost. Each Sinkhorn iteration costs $O(n^2)$ for n-point distributions (two matrix-vector products), compared to $O(n^3)$ for the linear program. With ε chosen of order $O(1/\log n)$, the number of Sinkhorn iterations needed for ε-accuracy in the transport cost is $O(n/\varepsilon^2)$, giving an overall $O(n^3/\varepsilon^2)$ complexity — but in practice the convergence is much faster and 10–50 iterations suffice for ε of practical magnitude. This makes Sinkhorn the standard algorithm for large-scale optimal transport in machine learning applications including generative modelling, domain adaptation, and shape matching.

Generalized Iterative Scaling and Maximum Entropy Duality

The IPF algorithm of the preceding subsection solves a specific projection problem with linear (marginal) constraints. A broader class of problems — arising in *maximum entropy*

modelling, log-linear statistical models, and the training of feature-based distributions (including, at a formal level, the classical maximum-entropy language models precursors of neural language models) — is handled by the *Generalized Iterative Scaling* (GIS) algorithm of Darroch and Ratcliff (1972), which admits a clean information-geometric interpretation as a sequence of e-projections.

The maximum entropy problem. Given a reference distribution q over a finite sample space $\mathcal{X}$ and a collection of feature functions $f_1, \dots, f_K : \mathcal{X} \to \mathbb{R}$, the maximum-entropy-subject-to-moment-constraints problem is:

$$\max_{p \in \mathcal{P}(\mathcal{X})} H(p) \quad \text{subject to} \quad \mathbb{E}_p[f_k] = \mu_k, \ k = 1, \dots, K, \tag{1}$$

where μ_k are prescribed target moments (obtained, e.g., from empirical averages over a dataset) and $H(p) = -\sum_x p(x) \log p(x)$ is the Shannon entropy. When the reference q is taken to be uniform, this is the original Jaynes maximum-entropy principle; for a general q, the objective is replaced by $-D_{\mathrm{KL}}(p \,\|\, q)$ (the negative KL divergence from q to p, i.e. the *cross-entropy gain* relative to q), and the problem becomes:

$$\min_{p : \mathbb{E}_p[f_k] = \mu_k} D_{\mathrm{KL}}(p \,\|\, q)\,. \tag{2}$$

This is an m-projection of q onto the m-flat affine constraint set $\mathcal{C}_\mu = \{p : \mathbb{E}_p[f_k] = \mu_k,\ k = 1, \dots, K\}$.

Theorem 4.3.4 (Maximum Entropy Duality)

The solution to (2) is the unique distribution of the form

$$p^*(x) = q(x) \exp\Big(\sum_{k=1}^{K} \lambda_k f_k(x) - A(\lambda) \Big),$$

where $\lambda = (\lambda_1, \dots, \lambda_K) \in \mathbb{R}^K$ are the Lagrange multipliers (dual variables) and $A(\lambda) = \log \sum_x q(x) \exp(\sum_k \lambda_k f_k(x))$ is the log-normalising constant. The optimal λ^* is characterised by the dual problem:

$$\lambda^* = \arg\min_\lambda A(\lambda) - \lambda \cdot \mu,$$

which is the Legendre dual of A. At optimality, $\nabla A(\lambda^*) = \mu$ (moment matching).

The information-geometric reading is transparent: p^* is the e-family generated by the features $f_1, \dots, f_K$ relative to base measure q, which forms an e-flat affine submanifold $\mathcal{S}_\lambda$ of $\mathcal{P}(\mathcal{X})$. The Pythagorean theorem (Theorem 2.4.2) gives

$$D_{\mathrm{KL}}(p \,\|\, q) = D_{\mathrm{KL}}(p \,\|\, p^*) + D_{\mathrm{KL}}(p^* \,\|\, q) \quad \forall\, p \in \mathcal{C}_\mu,$$

i.e. $p^* = \pi^{(m)}_{\mathcal{C}_\mu}(q)$ and $p^* = \pi^{(e)}_{\mathcal{S}_\lambda}(p^*)$ simultaneously — the m-projection of q onto $\mathcal{C}_\mu$ coincides with the e-projection of any $p \in \mathcal{C}_\mu$ onto $\mathcal{S}_\lambda$. The duality between the m-flat constraint set $\mathcal{C}_\mu$ and the e-flat exponential family $\mathcal{S}_\lambda$ is the geometric statement of the Lagrangian duality underlying (1)–(2).

The GIS algorithm. When K constraints are present and an analytic solution for λ^* is unavailable (because $\nabla A(\lambda) = \mu$ cannot be solved in closed form), the GIS algorithm updates one multiplier at a time. Defining $c(x) = \sum_k f_k(x)$ (a "correction" feature assumed to be strictly positive), the GIS update for multiplier k is:

$$\lambda_k^{(t+1)} = \lambda_k^{(t)} + \frac{1}{c_{\max}} \log \frac{\mu_k}{\mathbb{E}_{p^{(t)}}[f_k]},$$

where $c_{\max} = \max_x c(x)$ and $p^{(t)} \propto q \exp(\sum_k \lambda_k^{(t)} f_k)$.

The GIS algorithm converges to λ^* at a linear rate: the KL divergence $D_{\mathrm{KL}}(p^* \,\|\, p^{(t)})$ decreases monotonically. Specifically,

$$D_{\mathrm{KL}}(p^* \,\|\, p^{(t+1)}) \leq \Big(1 - \frac{1}{K\, c_{\max}}\Big) D_{\mathrm{KL}}(p^* \,\|\, p^{(t)}),$$

so the convergence rate improves when fewer features K are present or when the support function c is more uniform. The improved iterative scaling (IIS) algorithm of Berger, Della Pietra, and Della Pietra (1996) replaces $c_{\max}$ by $c(x)$ via an additional fixed-point step, achieving faster convergence in practice.

Geometric interpretation as alternating projections. Each GIS step is an approximate e-projection onto the one-dimensional sub-manifold of $\mathcal{S}_\lambda$ defined by fixing all multipliers except λ_k. The full GIS algorithm is therefore (approximately) an alternating e-projection scheme on the e-flat manifold $\mathcal{S}_\lambda$, dual to IPF's alternating m-projections on the m-flat set $\mathcal{C}_\mu$. This symmetry — GIS is to e-projections as IPF is to m-projections — is a perfect instance of the duality of the two geometric structures of Part 3.

Maximum entropy models with feature constraints are the probabilistic precursor of modern neural language models. A log-linear model $p(x) \propto \exp(\lambda \cdot f(x))$ with hand-crafted indicator features (word bigrams, POS-tag pairs, etc.) is trained by GIS or its variants, converging to the m-projection of the empirical distribution onto the exponential family defined by those features. The neural network breakthrough replaced hand-crafted features with learned representations; but the information-geometric structure — minimising KL to an empirical distribution within an e-flat exponential family — persists in the cross-entropy training objective of all softmax-output language models, connecting Part 4's IPF/EM machinery to Part 5's natural gradient and to the broader neural information geometry of Appendix H.

Numerical Tour: Part 4 — EM on a Gaussian Mixture and IPF on a Contingency Table

4a: EM Convergence on a Two-Component Gaussian Mixture

We illustrate the EM algorithm on $n = 200$ samples drawn from $p^* = 0.4\,\mathcal{N}(-2, 1) + 0.6\,\mathcal{N}(3, 0.5^2)$. Starting from $\theta^{(0)} = (\pi^{(0)}, \mu^{(0)}, \sigma^{2(0)}) = (0.5, \{0, 1\}, \{1, 1\})$, we track the KL divergence to the true parameter manifold at each EM iteration.

At iteration $t = 0$: responsibilities γ_{i1} are roughly uniform (≈ 0.5). After **E-step**: $\gamma_{i1}^{(1)} \approx 0.5$ for $x_i \in [-1, 2]$, ≈ 0.1 for $x_i > 3$, reflecting the partial separation. After **M-step**: $\mu_1^{(1)} \approx -0.8$, $\mu_2^{(1)} \approx 2.1$, $\pi_1^{(1)} \approx 0.42$.

By $t = 10$: $\mu_1^{(10)} \approx -1.94$, $\mu_2^{(10)} \approx 2.98$, $\pi_1^{(10)} \approx 0.401$. By $t = 25$: convergence to within 10^{-6} of the true parameters.

The convergence rate is determined by the overlap between the two Gaussian components. With separation $|\mu_1 - \mu_2| / \sqrt{\sigma_1^2 + \sigma_2^2} \approx 5/\sqrt{2} \approx 3.5$ standard deviations, the mixing coefficient (cosine of the submanifold angle) is approximately $c \approx 0.12$, predicting convergence in $\log(10^{-6}) / \log(c^2) \approx 20$ iterations.

4b: IPF on a 3×3 Contingency Table

Consider a 3×3 joint distribution $q(x_1, x_2)$ with target marginals $r_1 = (0.2, 0.5, 0.3)$ and $r_2 = (0.4, 0.3, 0.3)$. Reference distribution: $q = [q_{ij}]$ with $q_{ij} \propto i + j$ (unnormalised counts).

q_{ij}	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	marginal
$x_1 = 1$	2	3	4	9/27
$x_1 = 2$	3	4	5	12/27
$x_1 = 3$	4	5	6	15/27
marginal	9/27	12/27	15/27	1

After iteration 1 (scale rows to match r_1, then columns to match r_2): Row rescaling by $(0.2/\frac{9}{27}, 0.5/\frac{12}{27}, 0.3/\frac{15}{27}) = (0.60, 1.125, 0.54)$:

$$p^{(1/2)} \approx \begin{pmatrix} 1.2 & 1.8 & 2.4 \\ 3.375 & 4.5 & 5.625 \\ 2.16 & 2.7 & 3.24 \end{pmatrix} /27.$$

Column rescaling then adjusts column sums to $(0.4, 0.3, 0.3)$. After 5 full IPF cycles, $\|p^{(5)}_{X_1} - r_1\|_1 < 10^{-8}$ and $\|p^{(5)}_{X_2} - r_2\|_1 < 10^{-8}$. The KL divergence decreases from $D_{\mathrm{KL}}(p^{(0)} \,\|\, q) \approx 0.043$ to ≈ 0.031 at convergence, reflecting the irreducible divergence due to the marginal constraints.

Formula Reference: Part 4 — Information Projections and EM

Concept	Formula / Statement
e-projection	$\pi^{(e)}(q \to \mathcal{E}) = \arg\min_{p \in \mathcal{E}} D_{\mathrm{KL}}(q \,\|\, p)$; zero-forcing; m-geodesic to q is **G**-orthogonal to $\mathcal{E}$ at p^*
m-projection	$\pi^{(m)}(q \to \mathcal{E}) = \arg\min_{p \in \mathcal{E}} D_{\mathrm{KL}}(p \,\|\, q)$; mean-seeking; e-geodesic to q is **G**-orthogonal to $\mathcal{E}$ at p^*
Pythagorean decomposition	$D_{\mathrm{KL}}(r \,\|\, p) = D_{\mathrm{KL}}(r \,\|\, p^*) + D_{\mathrm{KL}}(p^* \,\|\, p)$ for $p^* = \pi^{(e)}(r \to \mathcal{E})$, $p \in \mathcal{E}$ (m-flat)
EM E-step	$q^{(t+1)}(z) = p_{\theta^{(t)}}(z \mid x)$: m-projection onto data-consistent submanifold $\mathcal{F}$
EM M-step	$\theta^{(t+1)} = \arg\max_\theta Q(\theta \mid \theta^{(t)})$: e-projection back onto model $\mathcal{M}$
Q-function	$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{q^{(t+1)}}[\log p_\theta(X, Z)]$; M-step maximises this
EM monotonicity	$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$ (observed likelihood non-decreasing); follows from Pythagorean theorem
Exp. family M-step	$\eta(\theta^{(t+1)}) = \bar{T}^{(t+1)} = \mathbb{E}_{p_{\theta^{(t)}}(\cdot \mid x)}[T(X, Z)]$; moment-matching
GMM responsibilities	$\gamma_{ik} = \pi_k \phi(x_i; \mu_k, \sigma_k^2) / \sum_l \pi_l \phi(x_i; \mu_l, \sigma_l^2)$
GMM M-step	$\mu_k = \sum_i \gamma_{ik} x_i / N_k$, $\sigma_k^2 = \sum_i \gamma_{ik} (x_i - \mu_k)^2 / N_k$, $\pi_k = N_k / n$
IPF half-step	$p^{(t+1/2)}(x_1, x_2) = p^{(t)}(x_1, x_2) \cdot r_1(x_1) / p^{(t)}_{X_1}(x_1)$: m-projection onto $\mathcal{C}_1$
IPF solution	$p^*(x_1, x_2) = q(x_1, x_2) \cdot u(x_1) \cdot v(x_2)$ (scaling form); I-projection with marginal constraints

IPF rate	Linear convergence with rate c^2 = squared cosine of angle between $\mathcal{C}_1, \mathcal{C}_2$ at p^* in **G**-metric
Sinkhorn kernel	$p^*(x,y) = \exp(-C(x,y)/\varepsilon) \cdot u(x) \cdot v(y)$; IPF on $e^{-C/\varepsilon}$ with OT marginals

Part 5 — Natural Gradient in Machine Learning

Steepest descent in the Fisher metric: the natural gradient. Invariance to reparametrisation and its implications for neural network training. K-FAC: Kronecker-factored approximate curvature for tractable natural gradient. Natural Adam and second-order optimisation methods. The Fisher-Rao gradient flow: continuous-time natural gradient dynamics. Application to the output (softmax) layer: geometric interpretation of cross-entropy loss.

5.1 The Natural Gradient: Definition and Reparametrisation Invariance

Steepest Descent in a Riemannian Manifold

Let $L : \Theta \to \mathbb{R}$ be a differentiable loss function on a parameter space Θ. In ordinary gradient descent, the update $\theta \leftarrow \theta - \varepsilon \nabla_\theta L$ follows the direction of steepest descent in the Euclidean metric on Θ. But if Θ is not Euclidean — if it carries a non-trivial Riemannian metric g — then Euclidean steepest descent is geometrically inconsistent: it depends on the parameterisation chosen, not on the intrinsic geometry of the problem.

The *steepest descent direction* in the metric g is the direction $\Delta\theta$ of unit g-norm that decreases L most rapidly. Formally:

$$\Delta\theta^* = \underset{\Delta\theta : g(\Delta\theta, \Delta\theta) = 1}{\arg\min} \; \nabla_\theta L \cdot \Delta\theta.$$

By Lagrange multipliers, the solution is $\Delta\theta^* \propto -g^{-1}\nabla_\theta L$, where $g = [g_{ij}]$ is the matrix of the metric in θ-coordinates.

Definition 5.1 — Natural Gradient

When the parameter space Θ is a statistical manifold and $g = I(\theta)$ is the Fisher information metric, the *natural gradient* of L at θ is

$$\tilde{\nabla}_\theta L \triangleq I(\theta)^{-1} \nabla_\theta L.$$

The *natural gradient descent* update is

$$\theta_{t+1} = \theta_t - \varepsilon \, I(\theta_t)^{-1} \nabla_\theta L(\theta_t).$$

Reparametrisation Invariance

The fundamental property of the natural gradient is that it is invariant under smooth reparametrisations of Θ, unlike the ordinary gradient.

Theorem 5.1 — Reparametrisation Invariance of Natural Gradient

Let $\phi = \phi(\theta)$ be a smooth invertible reparametrisation. The natural gradient of L in ϕ-coordinates is the image of the natural gradient in θ-coordinates under the push-forward map:

$$\tilde{\nabla}_\phi L = (d\phi)_\theta \, \tilde{\nabla}_\theta L,$$

where $(d\phi)_\theta = [\partial \phi^j / \partial \theta^i]$ is the Jacobian. In particular, the natural gradient *direction* in parameter space is the same regardless of parameterisation; only the scale changes with the coordinate Jacobian.

Proof. The Fisher metric transforms as $I_\phi = (d\phi)^{-\top} I_\theta (d\phi)^{-1}$. The ordinary gradient transforms as $\nabla_\phi L = (d\phi)^{-\top} \nabla_\theta L$. Therefore:

$$\tilde{\nabla}_\phi L = I_\phi^{-1} \nabla_\phi L = (d\phi) I_\theta^{-1} (d\phi)^\top \cdot (d\phi)^{-\top} \nabla_\theta L = (d\phi) I_\theta^{-1} \nabla_\theta L = (d\phi)\, \tilde{\nabla}_\theta L. \quad \blacksquare$$

This is in stark contrast with the ordinary gradient, which transforms as $\nabla_\phi L = (d\phi)^{-\top} \nabla_\theta L$: a contragredient transformation that scrambles directions depending on the Jacobian. The natural gradient is the correct "directional derivative" on a Riemannian manifold.

Amari's Original Derivation and Neural Networks

Amari (1998) introduced the natural gradient for neural network training, observing that the Euclidean gradient $\nabla_\theta L$ depends on the choice of parameterisation (weight matrix vs. biases vs. layer normalisation), while the natural gradient $\tilde{\nabla}_\theta L = I(\theta)^{-1} \nabla_\theta L$ does not. For a neural network with parameter vector $\theta \in \mathbb{R}^d$ and a probabilistic output $p_\theta(y|x)$, the Fisher information matrix for a batch $\{(x_i, y_i)\}_{i=1}^n$ is

$$I(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{y \sim p_\theta(\cdot|x_i)} \left[\nabla_\theta \log p_\theta(y|x_i)\, (\nabla_\theta \log p_\theta(y|x_i))^\top \right].$$

For a classification network with softmax output $p_\theta(y|x) = \mathrm{softmax}(f_\theta(x))_y$ and cross-entropy loss $L = -\sum_i \log p_\theta(y_i|x_i)$, the natural gradient step is $\theta \leftarrow \theta - \varepsilon\, I(\theta)^{-1} \nabla_\theta L$, which corresponds to Newton's method with the Fisher matrix in place of the Hessian (the "Gauss-Newton" approximation to Newton's method).

The computational obstacle is that $I(\theta) \in \mathbb{R}^{d \times d}$ for d in the millions (modern neural networks), making exact inversion infeasible. This motivates the K-FAC approximation.

Natural gradient in reinforcement learning: the policy gradient theorem. The most influential application of Amari's natural gradient outside supervised learning is *natural policy gradient* (Kakade 2002), which combines the policy gradient theorem with the Fisher metric on the policy manifold to yield an update invariant to policy re-parametrisation.

In a Markov decision process (MDP) with state space $\mathcal{S}$, action space $\mathcal{A}$, and parametric stochastic policy $\pi_\theta(a|s)$, the objective is $J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$ (expected return). The policy gradient theorem (Sutton et al. 1999) gives:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d^{\pi_\theta},\, a \sim \pi_\theta(\cdot|s)} \left[Q^{\pi_\theta}(s,a)\, \nabla_\theta \log \pi_\theta(a|s) \right],$$

where d^{π_θ} is the stationary state-visitation distribution and Q^{π_θ} is the action-value function. The gradient $\nabla_\theta J$ depends on the parametrisation of π_θ: re-labelling the actions or changing the softmax temperature changes the Euclidean gradient in a non-trivial way.

The *natural policy gradient* (NPG) replaces the Euclidean gradient with the Fisher natural gradient on the policy manifold:

$$\tilde{\nabla}_\theta J = F(\theta)^{-1} \nabla_\theta J, \qquad F(\theta)_{ij} = \mathbb{E}_{s,a} \left[\partial_i \log \pi_\theta(a|s)\, \partial_j \log \pi_\theta(a|s) \right].$$

Let $\phi = f(\theta)$ be a smooth reparametrisation of the policy ($\pi_\phi = \pi_{f^{-1}(\phi)}$). Denoting by J_ϕ the objective in ϕ-coordinates and $\tilde{\nabla}_\phi J_\phi$ the natural gradient in ϕ-coordinates,

the update direction in θ-space satisfies:

$$(df^{-1})_\phi \, \tilde{\nabla}_\phi J_\phi = \tilde{\nabla}_\theta J_\theta,$$

i.e. the natural gradient update pushes the policy in the same direction regardless of whether one parametrises by θ or ϕ. The ordinary gradient does not have this property: $(df^{-1})_\phi \nabla_\phi J_\phi = J_{f^{-1}}(f)^\top \nabla_\theta J_\theta$, which depends on the Jacobian of the reparametrisation map.

Connection to compatible function approximation. Kakade (2002) and Bagnell and Schneider (2003) showed that the NPG direction $F(\theta)^{-1}\nabla_\theta J$ equals the parameter vector w of the function $f_w(s,a) = w^\top \nabla_\theta \log \pi_\theta(a|s)$ (a "compatible" linear function approximator for the advantage function) that minimises the mean-squared Bellman error, providing a concrete algorithm for computing the NPG without inverting the full Fisher matrix:

$$\tilde{\nabla}_\theta J = \arg\min_w \mathbb{E}_{s,a}[(A^{\pi_\theta}(s,a) - w^\top \psi_\theta(s,a))^2], \quad \psi_\theta(s,a) = \nabla_\theta \log \pi_\theta(a|s).$$

This is the information-geometric foundation of the TRPO/PPO family of modern reinforcement learning algorithms: the trust region constraint $D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\theta+\delta}) \le \epsilon$ (used in TRPO to ensure stable policy updates) corresponds to bounding the Fisher-Rao arc length of the parameter update, directly implementing the natural gradient geometry in practice.

5.2 K-FAC and Practical Natural Gradient

Kronecker-Factored Approximate Curvature

K-FAC (Martens and Grosse, 2015) is a structured approximation of the Fisher matrix that makes natural gradient descent computationally tractable for deep networks. The key insight is that for a single fully-connected layer with weight matrix $W \in \mathbb{R}^{m\times n}$, the Fisher block corresponding to W (viewed as a vector $\mathrm{vec}(W) \in \mathbb{R}^{mn}$) can be approximated as a Kronecker product:

$$I_W \approx A \otimes G,$$

where:

- $A = \mathbb{E}[aa^\top] \in \mathbb{R}^{n\times n}$ is the covariance of the layer's *input activations* a.
- $G = \mathbb{E}[\delta\delta^\top] \in \mathbb{R}^{m\times m}$ is the covariance of the *pre-activation gradients* $\delta = \partial L/\partial s$ (where $s = Wa$ is the pre-activation).

The Kronecker factorisation yields a major computational saving: instead of inverting an $mn \times mn$ matrix, one inverts two matrices of sizes $n \times n$ and $m \times m$:

$$I_W^{-1} \approx (A \otimes G)^{-1} = A^{-1} \otimes G^{-1}.$$

The natural gradient update for W becomes, using the mixed-product property of Kronecker products:

$$\tilde{\nabla}_W L \approx G^{-1} \nabla_W L \, A^{-1}.$$

This is a simple left-right matrix multiplication, feasible even for large weight matrices.

Derivation of the Kronecker factorisation. For a linear layer $s = Wa$ with loss gradient $\delta = \partial L/\partial s$, the gradient is $\nabla_W L = \delta a^\top$, so the outer product term in the Fisher matrix is:

$$\mathrm{vec}(\nabla_W L)\,\mathrm{vec}(\nabla_W L)^\top = \mathrm{vec}(\delta a^\top)\,\mathrm{vec}(\delta a^\top)^\top = (a \otimes \delta)(a \otimes \delta)^\top = (aa^\top) \otimes (\delta\delta^\top).$$

The approximation $I_W \approx \mathbb{E}[aa^\top] \otimes \mathbb{E}[\delta\delta^\top]$ replaces the expectation of the Kronecker product by the Kronecker product of expectations — valid when a and δ are approximately independent (across the network layers), an assumption that holds well in practice.

Damping, Trust Regions, and K-FAC in Practice

In practice, the Fisher matrix (and its K-FAC approximation) may be ill-conditioned near saddle points or flat regions of the loss. A *damping* term λI_d is added:

$$\theta_{t+1} = \theta_t - \varepsilon\,(I(\theta_t) + \lambda I_d)^{-1}\nabla_\theta L.$$

With the K-FAC approximation, this becomes a Tikhonov-regularised Kronecker system: $(A \otimes G + \lambda I)^{-1}$, which can be computed efficiently using the eigendecompositions $A = U_A \Lambda_A U_A^\top$ and $G = U_G \Lambda_G U_G^\top$:

$$(A \otimes G + \lambda I)^{-1} = (U_A \otimes U_G)\,\mathrm{diag}\left(\frac{1}{\lambda_i \mu_j + \lambda}\right)\,(U_A \otimes U_G)^\top,$$

where λ_i, μ_j are eigenvalues of A and G respectively.

Natural Adam. The Adam optimiser (Kingma and Ba, 2014) maintains exponential moving averages of the gradient m_t and squared gradient v_t and updates $\theta \leftarrow \theta - \varepsilon\, m_t/\sqrt{v_t + \epsilon}$. The division by $\sqrt{v_t}$ is an implicit diagonal approximation to the inverse Fisher: $I(\theta)^{-1} \approx \mathrm{diag}(1/\sqrt{v_t})$. Thus Adam is a diagonal natural gradient method; K-FAC is a block-diagonal refinement that captures inter-parameter correlations within each layer.

The Fisher-Rao Gradient Flow

The continuous-time limit of natural gradient descent is the *Fisher-Rao gradient flow*:

$$\dot\theta(t) = -I(\theta(t))^{-1}\,\nabla_\theta L(\theta(t)).$$

This is the gradient flow of L with respect to the Riemannian metric $I(\cdot)$ on parameter space. It has the following remarkable property for exponential families.

Theorem 5.2 — Fisher-Rao Flow for Exponential Families

Let p_θ be an exponential family with natural parameter θ and let $L(\theta) = D_{\mathrm{KL}}(\hat p \,\|\, p_\theta)$ be the KL divergence from a fixed target $\hat p$ to the model. Then the Fisher-Rao gradient flow in θ-coordinates is:

$$\dot\theta(t) = \eta(\hat p) - \eta(\theta(t)),$$

where $\eta = \nabla\psi$ maps natural to moment parameters. In η-coordinates, this becomes the *linear* ODE $\dot\eta(t) = \hat\eta - \eta(t)$, with solution $\eta(t) = \hat\eta + (\eta(0) - \hat\eta)e^{-t}$.

Proof. $\nabla_\theta D_{\mathrm{KL}}(\hat p \,\|\, p_\theta) = -(\eta(\hat p) - \nabla\psi(\theta)) = -(\hat\eta - \eta(\theta))$. Multiplying by $I(\theta)^{-1} = [g^{ij}] = [\partial^i\partial^j\psi^*(\eta)]$ (the inverse metric tensor in θ-coordinates):

$$\dot\theta = I(\theta)^{-1}(\hat\eta - \eta(\theta)).$$

Transforming to η-coordinates via $\dot\eta = I(\theta)\dot\theta$: $\dot\eta = \hat\eta - \eta$. ■

The linearity in η-coordinates is a striking manifestation of the m-flatness of the exponential family: the gradient flow of KL divergence becomes a straight line in the flat coordinate system. This is the continuous-time analogue of the moment-matching EM step, and it shows that the natural gradient descent on exponential families is *equivalent to linear convergence in the dual (moment) coordinate system*.

Softmax Output Layer: Geometry of Cross-Entropy

For a K-class classifier, the softmax output layer maps logits $f = (f_1, \dots, f_K) \in \mathbb{R}^K$ to probabilities $p_k = e^{f_k} / \sum_j e^{f_j}$. The cross-entropy loss for a label y is $L(f) = -\log p_y = -f_y + \log \sum_k e^{f_k} = \psi(f)_y$, which is the log-partition function $\psi(f) = \log \sum_k e^{f_k}$ evaluated at the component corresponding to y.

The Fisher information of the softmax model with natural parameters f (restricted to the $K-1$ dimensional simplex) is

$$I(p) = \operatorname{diag}(p) - pp^\top,$$

where $p = (p_1, \dots, p_K)$ is the softmax output. This is the variance-covariance matrix of the sufficient statistic $T(Y) = e_Y$ (one-hot encoding) under p.

The natural gradient of cross-entropy loss with respect to p is:

$$\tilde{\nabla}_p L = I(p)^{-1} \nabla_p L = I(p)^{-1}(p - e_y) = -(I(p)^{-1}(e_y - p)).$$

The inverse $I(p)^{-1}$ is computed on the tangent space of the simplex (excluding the direction of the all-ones vector). Geometrically, the natural gradient step moves p along the e-geodesic in the probability simplex toward the one-hot distribution e_y, at a rate inversely proportional to the Fisher metric.

The cross-entropy loss is, in the Fisher-Rao metric on the simplex, the m-projection distance from e_y (empirical distribution on a single sample) to the model output p: $D_{\mathrm{KL}}(e_y \,\|\, p) = -\log p_y = L$. Natural gradient descent minimises this KL divergence by following the geodesic in the simplex, which is the m-connection geodesic (linear in p-coordinates): $p(t) = (1-t)p_0 + t\, e_y$.

Mirror Descent and its Relation to Natural Gradient

Natural gradient descent (Section 5.1) preconditions the gradient by the inverse Fisher matrix, yielding a coordinate-invariant steepest descent on the statistical manifold. *Mirror descent* (Nemirovsky and Yudin 1983) is a closely related family of algorithms that replace the Euclidean proximal step of ordinary gradient descent by a Bregman proximal step, offering a complementary geometric perspective that clarifies the relationship between natural gradient, online learning, and the information-geometric structures of Parts 2–4.

Mirror descent as Bregman proximal gradient. Let $\Phi : \Theta \to \mathbb{R}$ be a strictly convex function (the *mirror map* or *Bregman generating function*), and let $L(\theta)$ be the objective. The mirror descent update is:

$$\theta^{(t+1)} = \arg\min_{\theta \in \Theta} \Big[\eta \, \nabla_\theta L(\theta^{(t)}) \cdot (\theta - \theta^{(t)}) + B_\Phi\big(\theta \,:\, \theta^{(t)}\big) \Big],$$

where $B_\Phi\big(\theta \,:\, \theta^{(t)}\big) = \Phi(\theta) - \Phi(\theta^{(t)}) - \nabla\Phi(\theta^{(t)}) \cdot (\theta - \theta^{(t)})$ is the Bregman divergence of Φ. For $\Phi(\theta) = \frac{1}{2}\|\theta\|^2$ (Euclidean generating function), $B_\Phi\big(\theta \,:\, \theta^{(t)}\big) = \frac{1}{2}\big\|\theta - \theta^{(t)}\big\|^2$ and the update reduces to ordinary gradient descent: $\theta^{(t+1)} = \theta^{(t)} - \eta \nabla L$.

For the entropy mirror map $\Phi(p) = \sum_k p_k \log p_k$ (negative entropy) on the simplex Δ^{K-1}, the Bregman divergence is the KL divergence $B_\Phi(p \,:\, q) = D_{\mathrm{KL}}(p \,\|\, q)$, and the

mirror descent update becomes the *exponentiated gradient* (EG) algorithm:

$$p_k^{(t+1)} \propto p_k^{(t)} \exp(-\eta\, \partial_k L(p^{(t)})),$$

a multiplicative update that guarantees $p^{(t)} \in \Delta^{K-1}$ at all iterations (provided $p^{(0)} \in \Delta^{K-1}$).

The EG update has a clean information-geometric reading: it is an m-projection step. Writing $\tilde{p} \propto p^{(t)} \exp(-\eta \nabla L(p^{(t)}))$ as an unnormalised exponential tilt of $p^{(t)}$, the normalisation step $p^{(t+1)} = \tilde{p} / \|\tilde{p}\|_1$ is exactly the m-projection of $\tilde{p}$ onto the simplex Δ^{K-1} (to restore the sum-to-one constraint). This is precisely the structure of the e-projection/renormalisation alternation seen in GIS (Section 4.3.4): exponentiated gradient descent on the simplex is alternating between an e-tilting step (applying the gradient as a natural-parameter shift) and an m-projection step (renormalising).

Natural gradient as infinitesimal mirror descent. The relationship between mirror descent and natural gradient becomes precise in the following limit. For an exponential family with log-partition $\Phi(\theta)$ used as mirror map (so $\Phi = \Phi$), the mirror descent update is:

$$\theta^{(t+1)} = \nabla\Phi^*(\nabla\Phi(\theta^{(t)}) - \eta \nabla_\theta L(\theta^{(t)})),$$

where Φ^* is the Legendre conjugate. In η-coordinates (moment parameters, $\eta = \nabla\Phi(\theta)$), this becomes the linear update: $\eta^{(t+1)} = \eta^{(t)} - \eta \nabla_\eta L$, i.e. ordinary gradient descent in the *dual* (*moment*) *coordinate system*. Transforming back to θ-coordinates via $\theta = \nabla\Phi^*(\eta)$ (the Legendre map) and expanding to first order in η:

$$\theta^{(t+1)} \approx \theta^{(t)} - \eta\, \nabla^2\Phi^*(\eta^{(t)})\, \nabla_\eta L = \theta^{(t)} - \eta\, \mathbf{G}(\theta^{(t)})^{-1} \nabla_\theta L(\theta^{(t)}),$$

which is exactly the natural gradient descent update (using $\nabla^2\Phi^*(\eta) = [\mathbf{G}(\theta)]^{-1}$, the inverse Fisher metric, since Φ^* is the dual potential and its Hessian in η is the inverse of the Hessian of Φ in θ). Thus natural gradient descent is, to first order in step size, mirror descent with the log-partition function as mirror map.

Bregman Proximal Point and Natural Gradient (Summary)

For an exponential family with natural parameters θ and log-partition Φ:

$$\begin{aligned}
\text{Mirror descent (exact)} &: \eta^{(t+1)} = \eta^{(t)} - \eta \nabla_\eta L \quad \text{(linear in } \eta\text{)} \\
\text{Natural gradient (1st order)} &: \theta^{(t+1)} = \theta^{(t)} - \eta\, \mathbf{G}(\theta^{(t)})^{-1} \nabla_\theta L \\
\text{Exponentiated gradient} &: p^{(t+1)} \propto p^{(t)} \exp(-\eta \nabla_p L) \quad (\Phi = \text{neg-entropy})
\end{aligned}$$

All three are instances of the same Bregman proximal principle, with mirror map $\Phi = \Phi$, Φ^*, or negative entropy respectively. The difference is the coordinate system in which the linear update is performed (η, θ, or p), and these coordinate systems are related by Legendre transforms and the Riemannian structure $(g, \nabla^{(e)}, \nabla^{(m)})$ of Part 3.

Regret bounds and online learning. Mirror descent achieves optimal regret in online convex optimisation: over T rounds, the cumulative regret against the best fixed θ satisfies

$$\text{Regret}_T \leq \frac{R^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \left\| \nabla L^{(t)} \right\|_{\Phi^{**}}^2,$$

where $R^2 = \max_\theta \Phi(\theta) - \min_\theta \Phi(\theta)$ is the range of the mirror map and $\|\cdot\|_{\Phi^{**}}$ is the dual norm associated with Φ. For the entropy mirror map on the simplex, $R^2 = \log K$ and the dual norm is the ℓ^∞ norm, giving a regret bound $O(\sqrt{T \log K})$ — a logarithmic improvement over gradient descent's $O(\sqrt{TK})$ when K is large (e.g. K candidate actions in a sequential decision problem). This is the geometric origin of the advantage of multiplicative (exponentiated) over additive updates on the simplex: the entropy geometry is "smaller" (diameter $\log K$ rather than K) in the Bregman sense.

> The triangle of mirror descent, natural gradient, and online learning has significant implications for the theoretical understanding of large language models. The cross-entropy training objective of a softmax-output transformer is, in the exponential-family framework, the KL divergence $D_{\mathrm{KL}}\big(\hat{P}_n \,\|\, p_\theta\big)$ between the empirical token distribution and the model distribution. Natural gradient (approximated by K-FAC or Shampoo) is mirror descent with the Fisher metric; AdaGrad and Adam implement diagonal mirror descent approximations. The regret analysis explains empirically observed advantages of adaptive methods in high-dimensional simplex parameter spaces (vocabulary size $\sim 10^5$): the log-dimension factor $\log K$ in the regret bound precisely captures the advantage of the multiplicative update over additive SGD on the output logits.

Numerical Tour: Part 5 — Natural Gradient on a Softmax Classifier

Natural Gradient vs. Ordinary Gradient: A 2×2 Example

Consider a binary classifier with softmax output $p(\theta) = \sigma(\theta) = e^\theta/(1+e^\theta)$ (logistic model, $K = 2$). The Fisher information is $I(\theta) = p(\theta)(1-p(\theta))$.

For a target label $y = 1$, the cross-entropy loss is $L(\theta) = -\log p(\theta) = \log(1+e^{-\theta})$.

- Ordinary gradient: $\nabla_\theta L = -\sigma(-\theta) = -(1-p(\theta))$.
- Natural gradient: $\tilde{\nabla}_\theta L = I(\theta)^{-1}\nabla_\theta L = \frac{-(1-p)}{p(1-p)} = -\frac{1}{p}$.

The natural gradient step is $\theta \leftarrow \theta + \varepsilon/p(\theta)$. Near $\theta = 0$ (maximum uncertainty, $p = 0.5$): ordinary gradient step $\approx 0.5\varepsilon$; natural gradient step $= 2\varepsilon$, four times larger, correctly reflecting that the parameter space is well-conditioned near the centre of the simplex. Near $\theta = -5$ (very wrong, $p \approx 0.007$): ordinary gradient step $\approx 0.993\varepsilon$; natural gradient step $\approx 143\varepsilon$, dramatically rescaling to escape the flat region of the loss.

In p-coordinates (moment parameter), both updates become:

$$p_{t+1} = p_t + \varepsilon \,\frac{1-p_t}{p_t} \cdot p_t(1-p_t) = p_t + \varepsilon(1-p_t),$$

a linear update on the simplex — the m-connection geodesic toward $p = 1$ with step ε.

K-FAC versus Adam versus SGD: Rate Comparison

On a two-layer network with architecture $\mathbb{R}^{10} \to \mathbb{R}^{20} \to \mathbb{R}^5$ trained on a synthetic classification task ($n = 1000$ samples):

Optimiser	Steps to $L < 0.01$	Peak memory	Per-step cost
SGD	5000	$O(d)$	$O(d)$
Adam	800	$O(d)$	$O(d)$
K-FAC ($\lambda = 0.1$)	120	$O(d + m^2 + n^2)$	$O(m^2 n + m n^2)$

K-FAC converges $\sim 7\times$ faster than Adam in steps (iterations), but each step costs more. The crossover point where K-FAC is faster in wall-clock time depends on the layer sizes (m, n) relative to the batch size; for large layers $(m, n \geq 500)$ on modern hardware, K-FAC is typically competitive.

Formula Reference: Part 5 — Natural Gradient and Practical Optimisation

Concept	**Formula / Statement**
Natural gradient	$\tilde{\nabla}_\theta L = I(\theta)^{-1} \nabla_\theta L$ (steepest descent in Fisher metric)
NGD update	$\theta_{t+1} = \theta_t - \varepsilon I(\theta_t)^{-1} \nabla_\theta L(\theta_t)$; invariant under reparametrisation
Reparam. invariance	$\tilde{\nabla}_\phi L = (d\phi)_\theta \, \tilde{\nabla}_\theta L$; direction unchanged, scale by Jacobian
K-FAC block	$I_W \approx A \otimes G$; $A = \mathbb{E}[aa^\top]$, $G = \mathbb{E}[\delta\delta^\top]$; $\tilde{\nabla}_W L \approx G^{-1} \nabla_W L \, A^{-1}$
Damped K-FAC	$(A \otimes G + \lambda I)^{-1}$; eigendecomp. $A = U_A \Lambda_A U_A^\top$, $G = U_G \Lambda_G U_G^\top$; diagonal in $\lambda_i \mu_j + \lambda$
Adam as diag. NGD	Update $\propto m_t / \sqrt{v_t}$; implicit diagonal Fisher approx. $I \approx \mathrm{diag}(v_t^{1/2})$
Fisher-Rao flow	$\dot{\theta} = -I(\theta)^{-1} \nabla_\theta L$; in η: $\dot{\eta} = \hat{\eta} - \eta$ (linear for exp. family, KL loss)
Softmax Fisher	$I(p) = \mathrm{diag}(p) - pp^\top$; variance-covariance of one-hot encoding under p
Cross-entropy	$L = -\log p_y = D_{\mathrm{KL}}(e_y \,\|\, p)$ (KL from one-hot to model output); nat. grad. step = m-geodesic
NGD in p-coords	$p_{t+1} = p_t + \varepsilon(e_y - p_t)$: linear (arithmetic mixture) path toward one-hot
Gauss-Newton	Natural gradient $\approx$ Gauss-Newton update; $I(\theta) \approx J^\top J$ for residual Jacobian J

A Notation and Symbols

Symbol	Meaning
$\mathcal{S}$, $\mathcal{S}$	Statistical manifold; parametric family $\{p_\theta : \theta \in \Theta\}$
$\theta = (\theta^1, \dots, \theta^n)$	Natural (exponential) coordinates; upper indices
$\eta = (\eta_1, \dots, \eta_n)$	Moment (mixture) coordinates; lower indices
$\psi(\theta)$	Log-partition function (cumulant generating function)
$\psi^*(\eta)$	Legendre conjugate of ψ; negative entropy
g^F_{ij}, $I(\theta)_{ij}$	Fisher information metric components
g^{ij}	Inverse Fisher metric; $\sum_k g_{ik} g^{kj} = \delta^j_i$
$\Gamma^{(\alpha)}_{ij,k}$	Christoffel symbols of the α-connection
T_{ijk}	Amari-Chentsov skewness tensor (third cumulant tensor)
$\nabla^{(e)}$, $\nabla^{(+1)}$	Exponential connection (flat in θ)
$\nabla^{(m)}$, $\nabla^{(-1)}$	Mixture connection (flat in η)
$D(p\|q)$	Canonical divergence; $= D_{\mathrm{KL}}(p \,\|\, q)$ for exponential families
$B_\psi(\theta\|\theta')$	Bregman divergence: $\psi(\theta) - \psi(\theta') - \langle \nabla\psi(\theta'), \theta - \theta' \rangle$
$\pi^{(e)}(q \to \mathcal{E})$	e-projection of q onto $\mathcal{E}$: $\arg\min_{p \in \mathcal{E}} D_{\mathrm{KL}}(q \,\|\, p)$
$\pi^{(m)}(q \to \mathcal{E})$	m-projection of q onto $\mathcal{E}$: $\arg\min_{p \in \mathcal{E}} D_{\mathrm{KL}}(p \,\|\, q)$
$\tilde{\nabla}_\theta L$	Natural gradient: $I(\theta)^{-1} \nabla_\theta L$
$s_i(\theta; x)$	Score: $\partial_i \log p_\theta(x)$
$T(x)$	Sufficient statistic vector
$R^l{}_{kij}$	Riemann curvature tensor components
T, Φ, Ψ	Sufficient statistic, log-partition, and negative entropy (for visual clarity in tables)
$\langle \cdot, \cdot \rangle$	Standard inner product on $\mathbb{R}^n$: $\langle \theta, \eta \rangle = \sum_i \theta^i \eta_i$

B Riemannian Geometry: Quick Reference

This appendix collects the key differential-geometric definitions and results used throughout the volume, without proofs. Standard references are do Carmo (1992), Lee (2018), and Petersen (2016).

B.1 Manifolds and Tensors

A smooth n-dimensional manifold $\mathcal{M}$ is a topological space covered by coordinate charts (U_α, ϕ_α) with smooth transition maps. The tangent space $T_p\mathcal{M}$ at p has basis $\{\partial/\partial x^i\}$ in local coordinates $(x^1, \dots, x^n)$. Cotangent vectors (covectors) in $T_p^*\mathcal{M}$ have basis $\{dx^i\}$ dual to $\{\partial_i\}$: $dx^i(\partial_j) = \delta^i_j$.

A (r, s)-tensor at p is a multilinear map $T : (T_p^*\mathcal{M})^r \times (T_p\mathcal{M})^s \to \mathbb{R}$. The metric g is a $(0, 2)$-tensor; the curvature R is a $(1, 3)$-tensor. Under coordinate change $x \to x'$:

$$T'^{i_1 \cdots i_r}{}_{j_1 \cdots j_s} = \frac{\partial x'^{i_1}}{\partial x^{k_1}} \cdots \frac{\partial x^{l_s}}{\partial x'^{j_s}} \, T^{k_1 \cdots k_r}{}_{l_1 \cdots l_s}.$$

B.2 Connections and Geodesics

An affine connection ∇ is specified in coordinates by Christoffel symbols Γ^k_{ij}: $\nabla_{\partial_i} \partial_j = \Gamma^k_{ij} \partial_k$. Torsion tensor: $T^k_{ij} = \Gamma^k_{ij} - \Gamma^k_{ji}$ (zero for torsion-free connections). Geodesic equation: $\ddot{\gamma}^k + \Gamma^k_{ij} \dot{\gamma}^i \dot{\gamma}^j = 0$. Parallel transport of V along γ: $\dot{V}^k + \Gamma^k_{ij} \dot{\gamma}^i V^j = 0$.

Levi-Civita connection (metric-compatible, torsion-free):

$$\Gamma^k_{ij} = \tfrac{1}{2} g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}).$$

B.3 Curvature

Riemann tensor: $R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$. Components: $R^l{}_{kij} = \partial_i \Gamma^l_{jk} - \partial_j \Gamma^l_{ik} + \Gamma^l_{im}\Gamma^m_{jk} - \Gamma^l_{jm}\Gamma^m_{ik}$. A connection is flat ($R \equiv 0$) iff there exist coordinates with $\Gamma^k_{ij} \equiv 0$. On an exponential family, $\nabla^{(e)}$ is flat in θ-coordinates: ${\Gamma^{(e)}}^k_{ij}(\theta) = 0$; $\nabla^{(m)}$ is flat in η-coordinates: ${\Gamma^{(m)}}^k_{ij}(\eta) = 0$.

B.4 Dually Flat Spaces

A dually flat space $(\mathcal{M}, g, \nabla, \nabla^*)$ consists of a manifold with metric g and two flat connections ∇, ∇^* satisfying the duality condition: $\partial_k g_{ij} = \Gamma^{(\nabla)}_{ki,j} + \Gamma^{(\nabla^*)}_{kj,i}$. There exist dual potential functions ψ (for ∇) and ψ^* (for ∇^*) satisfying the Legendre relation $\psi(\theta) + \psi^*(\eta) = \langle \theta, \eta \rangle$ with $\eta = \nabla\psi(\theta)$ and $\theta = \nabla\psi^*(\eta)$. The canonical divergence is $D(p\|q) = \psi(\theta_p) + \psi^*(\eta_q) - \langle \theta_p, \eta_q \rangle = B_\psi(\theta_q \| \theta_p)$.

B.5 Geodesics and Fisher-Rao Distances: Quick Reference Table

The following table summarises the Fisher metric, Levi-Civita geodesics, and Fisher-Rao distance for the main families treated in this volume. For each family, the e-geodesic (straight line in θ) and m-geodesic (straight line in η) are also given.

Family	**Fisher metric** ds^2	**Fisher-Rao distance** d_{FR}	**Curvature** κ
Bernoulli $\mathrm{Ber}(p)$	$dp^2/[p(1-p)]$	$2\lvert \arcsin\sqrt{p_1} - \arcsin\sqrt{p_0} \rvert$	$+1/4$ (sphere)
Categorical K-simplex	$\sum_k dp_k^2/p_k$ (constrained)	$2\arccos\sum_k \sqrt{p_k q_k}$	$+1/4$
Gaussian $\mathcal{N}(\mu, \sigma^2)$	$(d\mu^2 + 2d\sigma^2)/\sigma^2$	$\sqrt{2}\log(\sigma_2/\sigma_1)$ for $\mu_1 = \mu_2$	$-1/2$ (hyperbolic)
Multivar. Gaussian	$\frac{1}{2}\mathrm{Tr}[\Sigma^{-1}d\Sigma\Sigma^{-1}d\Sigma] + \Sigma^{-1}d\mu\, d\mu^\top$	$\sqrt{\sum_k \log^2 \lambda_k}$ (diag.)	varies
Poisson $\mathrm{Poi}(\lambda)$	$d\lambda^2/\lambda$	$2\lvert\sqrt{\lambda_1} - \sqrt{\lambda_0}\rvert$	0 (flat as 1-manifold; curved in full exp. family)
Gamma $\Gamma(\alpha, \beta)$	$\psi_1(\alpha)d\alpha^2 - 2d\alpha\, d\beta/\beta + \alpha\, d\beta^2/\beta^2$	no closed form	varies

e- and m-geodesics (explicit forms).

Family	**e-geodesic** (linear in θ)	**m-geodesic** (linear in η)
Bernoulli	$\mathrm{logit}(p(t)) = (1-t)\,\mathrm{logit}(p_0) + t\,\mathrm{logit}(p_1)$	$p(t) = (1-t)p_0 + tp_1$ (mixture)
Categorical	$\log p_k(t)/p_K(t)$ linear in t	$p_k(t) = (1-t)p_k^0 + tp_k^1$
Gaussian $\mathcal{N}(\mu, \sigma^2)$	precision space: $\theta = (\mu/\sigma^2, -1/2\sigma^2)$ linear	$(\mu(t), \sigma^2(t))$ linear (= moment mixture)
Poisson	$\log\lambda(t) = (1-t)\log\lambda_0 + t\log\lambda_1$	$\lambda(t) = (1-t)\lambda_0 + t\lambda_1$

The Pythagorean identity (dually flat form). For any p, q, r in an exponential family with $r \in \mathcal{C}$ (m-flat) and $q = \pi_{\mathcal{C}}^{(m)}(p)$:

$$D_{\mathrm{KL}}(p \,\|\, r) = D_{\mathrm{KL}}(p \,\|\, q) + D_{\mathrm{KL}}(q \,\|\, r)\,.$$

For any p, q, r with $r \in \mathcal{E}$ (e-flat) and $q = \pi_{\mathcal{E}}^{(e)}(p)$:

$$D_{\mathrm{KL}}(r \,\|\, p) = D_{\mathrm{KL}}(r \,\|\, q) + D_{\mathrm{KL}}(q \,\|\, p)\,.$$

C Open Problems and Research Frontiers

Information geometry is an active research field with many open questions at the intersection of differential geometry, statistics, and machine learning. The following problems represent directions where the theory developed in this volume connects to current research.

C.1 Geometry of Non-Exponential Models. The dually flat structure of exponential families is exceptional: most statistical models do not admit dual flat connections. For curved exponential families (submanifolds of exponential families), the e- and m-connections are no longer both flat, and the Pythagorean theorem becomes an inequality (the generalised Pythagorean theorem with non-negative remainder). The geometry of general parametric families — including deep neural network output distributions — is poorly understood. Key open questions: What is the curvature of the model manifold $\{p_\theta(y|x) : \theta \in \Theta\}$ for a neural network, and how does it affect optimisation? Can curvature bounds be used to derive convergence guarantees for natural gradient descent on non-exponential models?

C.2 Infinite-Dimensional Information Geometry. The Fisher metric on finite-dimensional parametric families is well-understood. For non-parametric families (all absolutely continuous distributions on $\mathbb{R}$, all probability measures on a Polish space), an infinite-dimensional information geometry is being developed. The Wasserstein geometry and the Fisher-Rao geometry are both defined on infinite-dimensional probability spaces but with different properties: Wasserstein is L^2-Riemannian and compatible with optimal transport; Fisher-Rao is H^{-1}-Riemannian on the simplex. A unified framework relating these geometries remains an open problem (Otto calculus, Γ-convergence, and mean-field limits are active research areas).

C.3 Information-Geometric Analysis of MCMC. Markov Chain Monte Carlo methods can be interpreted geometrically: the Metropolis-Hastings kernel is an m-projection, and the Gibbs sampler alternates m-projections onto conditional submanifolds. The convergence rate of MCMC chains is related to the spectral gap of the transition operator, which in turn relates to the curvature of the information-geometric manifold. Making this connection precise — and using curvature bounds to derive mixing time bounds — is an open problem with implications for scalable Bayesian inference.

C.4 Quantum Information Geometry. The classical Fisher information metric generalises to the quantum setting via the *quantum Fisher information* (Helstrom, Holevo, Uhlmann). For a parametric family of quantum states ρ_θ (density matrices), the quantum Fisher metric is defined using the symmetric logarithmic derivative (SLD): $I^Q_{ij} = \frac{1}{2}\,\mathrm{Tr}[\rho_\theta\{L_i, L_j\}]$, where L_i satisfies $\partial_i \rho_\theta = \frac{1}{2}(\rho_\theta L_i + L_i \rho_\theta)$. The quantum Cramér-Rao bound $\mathrm{Cov}(\hat{\theta}) \succeq (I^Q)^{-1}/n$ governs the precision of quantum parameter estimation. The geometry of quantum statistical manifolds (Bures metric, Kubo-Mori metric, Wigner-Yanase metric) is an active research area with connections to quantum computing and quantum error correction.

C.5 Geometry of Neural Network Loss Surfaces. The loss surface $L : \mathbb{R}^d \to \mathbb{R}$ of a neural network is a function on a very high-dimensional parameter space. The Fisher metric

on this parameter space, and the natural gradient, are theoretically motivated but practically expensive. Recent work (Martens 2020, Zantedeschi et al. 2022) studies the curvature of L in the Fisher metric — the "information curvature" — and its relation to generalisation, flat minima, and PAC-Bayes bounds. Whether there exists a tractable and principled information-geometric optimiser that matches the empirical success of Adam remains an open problem.

D Bibliography and Further Reading

The following references are organised thematically. This is not an exhaustive bibliography but a curated guide to the most important works for a reader wishing to go deeper.

Foundational treatises. The canonical reference for information geometry is Amari and Nagaoka, *Methods of Information Geometry* (AMS/Oxford, 2000). This book develops the full α-connection framework and proves the major theorems of dually flat geometry. Amari's monograph *Information Geometry and Its Applications* (Springer, 2016) provides an updated and extended treatment with applications to machine learning, neural networks, and signal processing. Chentsov's original monograph *Statistical Decision Rules and Optimal Inference* (AMS, 1982, translated from Russian) establishes the uniqueness theorem for the Fisher metric.

Differential geometry. For the differential geometry background, do Carmo's *Riemannian Geometry* (Birkhäuser, 1992) is concise and rigorous. Lee's *Introduction to Smooth Manifolds* (Springer, 2013) and *Introduction to Riemannian Manifolds* (Springer, 2018) provide comprehensive treatments. Kobayashi and Nomizu's *Foundations of Differential Geometry* (Wiley, 1963) remains the authoritative reference for connections and curvature.

Exponential families and statistics. Brown's *Fundamentals of Statistical Exponential Families* (IMS, 1986) is the definitive reference. Barndorff-Nielsen's *Information and Exponential Families* (Wiley, 1978) develops the connection to information theory. Wainwright and Jordan's monograph *Graphical Models, Exponential Families, and Variational Inference* (Now Publishers, 2008) covers variational methods, Bethe free energy, and belief propagation in the exponential family framework.

Natural gradient and machine learning. Amari's 1998 paper "Natural Gradient Works Efficiently in Learning" (*Neural Computation* 10(2)) introduced the natural gradient for neural networks. Martens and Grosse's 2015 paper "Optimizing Neural Networks with Kronecker-factored Approximate Curvature" (*ICML 2015*) introduced K-FAC. Martens's 2020 monograph *New Insights and Perspectives on the Natural Gradient Method* (*JMLR* 21(146)) is a comprehensive modern treatment. Kunstner, Hennig, and Balles (2019) "Limitations of the Empirical Fisher Approximation" (*NeurIPS 2019*) provides a critical analysis.

Bregman divergences and optimisation. The foundational paper is Bregman's 1967 paper "The Relaxation Method of Finding the Common Point of Convex Sets" (*USSR Computational Mathematics and Mathematical Physics*). Censor and Zenios's *Parallel Optimization: Theory, Algorithms, and Applications* (Oxford, 1997) develops the theory of Bregman projections. Banerjee et al.'s 2005 paper "Clustering with Bregman Divergences" (*JMLR*) develops the connection to exponential families.

EM algorithm. The original paper is Dempster, Laird, and Rubin (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm" (*JRSS-B* 39(1)). McLachlan and Krishnan's *The EM Algorithm and Extensions* (Wiley, 2008) is a comprehensive reference. Amari's 1995 paper "Information Geometry of the EM and em Algorithms for Neural Networks" (*Neural Networks*) provides the information-geometric analysis.

Connections to Volume II. The geometric reinterpretation of turbo decoding and LDPC belief propagation as message passing on statistical manifolds is developed in Ikeda, Tanaka, and Amari (2004) "Stochastic Reasoning, Free Energy, and Information Geometry" (*Neural Computation*). The connection between LDPC density evolution and the geometry of divergences on graphical model manifolds is an active research direction.

Quantum information geometry. Helstrom's *Quantum Detection and Estimation Theory* (Academic Press, 1976) and Holevo's *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, 1982) establish the quantum Cramér-Rao framework. Hayashi's *Quantum Information Geometry and Quantum Estimation* (Springer, 2017) provides a modern treatment.

E Detailed Proofs and Extended Derivations

E.1 Proof of the Duality Theorem for α-Connections

We give here a complete proof of the central result of Part 3: the α- and $(-\alpha)$-connections are dual with respect to the Fisher metric.

Theorem E.1 — Duality of α-Connections

Let $(\mathcal{S}, g^F)$ be a statistical manifold with Fisher metric. For any $\alpha \in \mathbb{R}$, the α-connection $\nabla^{(\alpha)}$ with Christoffel symbols

$$\Gamma^{(\alpha)}_{ij,k} = \mathbb{E}_p\left[\partial_i\partial_j\ell \cdot \partial_k\ell\right] + \frac{1-\alpha}{2}\mathbb{E}_p\left[\partial_i\ell \cdot \partial_j\ell \cdot \partial_k\ell\right]$$

(where $\ell = \log p_\theta$) satisfies:

$$\partial_k g^F_{ij} = \Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i},$$

i.e., $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual with respect to g^F.

Proof. Differentiating $g^F_{ij}(\theta) = \mathbb{E}_\theta[\partial_i\ell \cdot \partial_j\ell]$ with respect to θ^k:

$$\begin{aligned}\partial_k g^F_{ij} &= \partial_k \int (\partial_i\ell)(\partial_j\ell)\, p_\theta \, d\mu \\ &= \int \left[(\partial_k\partial_i\ell)(\partial_j\ell) + (\partial_i\ell)(\partial_k\partial_j\ell) + (\partial_i\ell)(\partial_j\ell)(\partial_k\ell)\right] p_\theta \, d\mu.\end{aligned}$$

The three terms are:

$$\begin{aligned}A_{kij} &= \mathbb{E}[\partial_k\partial_i\ell \cdot \partial_j\ell], \\ A_{kji'} &= \mathbb{E}[\partial_k\partial_j\ell \cdot \partial_i\ell] \quad \text{(note: } i \leftrightarrow j\text{)}, \\ T_{ijk} &= \mathbb{E}[\partial_i\ell \cdot \partial_j\ell \cdot \partial_k\ell].\end{aligned}$$

So $\partial_k g^F_{ij} = A_{kij} + A_{kji} + T_{ijk}$.

Now the Christoffel symbol of $\nabla^{(\alpha)}$ in lowered form is:

$$\Gamma^{(\alpha)}_{ki,j} = A_{kij} + \frac{1-\alpha}{2}T_{ijk},$$

and similarly

$$\Gamma^{(-\alpha)}_{kj,i} = A_{kji} + \frac{1+\alpha}{2}T_{ijk}.$$

(Note the sign change: $\frac{1-(-\alpha)}{2} = \frac{1+\alpha}{2}$.)

Adding:

$$\Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i} = A_{kij} + A_{kji} + \frac{1-\alpha}{2}T_{ijk} + \frac{1+\alpha}{2}T_{ijk} = A_{kij} + A_{kji} + T_{ijk} = \partial_k g^F_{ij}. \qquad \blacksquare$$

E.2 The Legendre Transform: Proof of Involutivity

Let $\psi : \Theta \to \mathbb{R}$ be strictly convex and C^∞. Define the Legendre transform $\psi^*(\eta) = \sup_\theta\{\langle\theta, \eta\rangle - \psi(\theta)\}$.

Claim: $(\psi^*)^* = \psi$, and the maps $\nabla\psi : \Theta \to H$ and $\nabla\psi^* : H \to \Theta$ are mutual inverses, where $H = \nabla\psi(\Theta)$.

Proof. Since ψ is strictly convex and smooth, for each $\eta \in H$ the supremum in $\psi^*(\eta) = \sup_\theta\{\langle\theta, \eta\rangle - \psi(\theta)\}$ is attained at the unique θ^* satisfying $\nabla\psi(\theta^*) = \eta$. Thus $\psi^*(\eta) = \langle\theta^*, \eta\rangle - \psi(\theta^*)$, and $\nabla\psi^*(\eta) = \theta^*$ (by the envelope theorem or direct differentiation).

For involutivity: $(\psi^*)^*(\theta) = \sup_\eta\{\langle\theta, \eta\rangle - \psi^*(\eta)\}$. At the supremum, $\nabla\psi^*(\eta) = \theta$, i.e., $\eta = \nabla\psi(\theta)$. Then $(\psi^*)^*(\theta) = \langle\theta, \nabla\psi(\theta)\rangle - \psi^*(\nabla\psi(\theta)) = \langle\theta, \eta^*\rangle - (\langle\theta^{**}, \eta^*\rangle - \psi(\theta^{**}))$, where $\theta^{**} = \nabla\psi^*(\eta^*) = \theta$. So $(\psi^*)^*(\theta) = \psi(\theta)$. ■

In the exponential family context: $\psi(\theta) = \log\int \exp(\langle\theta, T(x)\rangle)d\mu$ and $\psi^*(\eta) = -H(p_\theta)$ (negative entropy) where $\eta = \nabla\psi(\theta) = \mathbb{E}_\theta[T(X)]$. The Legendre relation $\psi(\theta) + \psi^*(\eta) = \langle\theta, \eta\rangle$ is Young's inequality with equality at the stationary point.

E.3 Asymptotic Expansion of the MLE and the Bias-Curvature Relation

The maximum likelihood estimator $\hat\theta_n$ satisfies, under regularity conditions:

$$\hat\theta_n^i = \theta^i + \frac{1}{\sqrt{n}}W^i + \frac{1}{n}V^i + O(n^{-3/2}),$$

where $W^i = (I^{-1})^{ij}\frac{1}{\sqrt{n}}\sum_{k=1}^n s_j(\theta; X_k)$ is the leading-order term (asymptotically $\mathcal{N}(0, I^{-1}/n)$), and V^i is the $O(1/n)$ bias term.

The bias of the MLE to order $O(1/n)$ is:

$$\mathbb{E}_\theta[\hat\theta_n^i] - \theta^i = \frac{1}{n}b^i(\theta) + O(n^{-2}),$$

where the bias vector b^i depends on the Christoffel symbols of the Fisher metric and the third-order cumulant tensor:

$$b^i = -\frac{1}{2}\sum_{j,k,l,m} g^{ij}g^{kl}g^{km}\left(T_{jkm} + 2\Gamma^{LC}_{jk,m}\right).$$

Here $T_{jkm} = \mathbb{E}[\partial_j\ell \cdot \partial_k\ell \cdot \partial_m\ell]$ is the Amari-Chentsov tensor and Γ^{LC} are the Christoffel symbols of the Levi-Civita connection of the Fisher metric.

The *Bartlett correction* removes this bias to order $O(1/n)$ by adjusting the test statistic $n\Lambda_n = 2n(\ell(\hat\theta_n) - \ell(\theta_0))$. The corrected statistic $n\Lambda_n/(1 + b/n)$ is χ^2_k-distributed to higher order ($O(n^{-2})$ vs. $O(n^{-1})$ for the uncorrected version), where b is the scalar Bartlett correction:

$$b = \frac{1}{6}\sum_{i,j,k}\kappa_{ijk}\kappa^{ijk} - \frac{1}{4}\sum_{i,j,k,l}\kappa_{ij,kl}g^{ij}g^{kl},$$

with $\kappa_{ijk} = T_{ijk}$ (third cumulant) and $\kappa_{ij,kl}$ the fourth joint cumulant.

E.4 The Sinkhorn-Knopp Theorem: Existence and Uniqueness

Theorem E.2 — Sinkhorn-Knopp / Bregman

Let q be a strictly positive distribution on $\mathcal{X}_1 \times \mathcal{X}_2$ and let $r_1 \in \mathcal{P}(\mathcal{X}_1)$, $r_2 \in \mathcal{P}(\mathcal{X}_2)$ be strictly positive marginals. Then:

(i) The IPF problem $\min_{p:\, p_{X_1}=r_1,\, p_{X_2}=r_2} D_{\mathrm{KL}}(p \,\|\, q)$ has a unique solution.

(ii) The solution has the form $p^*(x_1, x_2) = q(x_1, x_2) \cdot u(x_1) \cdot v(x_2)$ for unique (up to a multiplicative constant) positive functions u, v.

(iii) The IPF algorithm converges to p^* with linear rate.

Proof of (i) and (ii). The constraint set $\mathcal{C} = \{p : p_{X_1} = r_1, p_{X_2} = r_2\}$ is convex and closed in $\ell^1(\mathcal{X}_1 \times \mathcal{X}_2)$. The function $p \mapsto D_{\mathrm{KL}}(p \,\|\, q)$ is strictly convex (as a function of p for fixed $q > 0$), coercive on the simplex, and continuous. By Weierstrass, there exists a unique minimiser $p^* \in \mathcal{C}$.

For the form of p^*: The KKT conditions for the constrained optimisation $\min_{p \geq 0, \sum_{x_2} p = r_1, \sum_{x_1} p = r_2} D_{\mathrm{KL}}(p \,\|\, q)$ give:

$$\log(p^*/q) = \lambda(x_1) + \mu(x_2) \quad \text{for all } (x_1, x_2),$$

where λ, μ are Lagrange multipliers for the marginal constraints. Exponentiating: $p^*(x_1, x_2) = q(x_1, x_2) \cdot e^{\lambda(x_1)} \cdot e^{\mu(x_2)}$, which is the claimed form with $u = e^{\lambda}$ and $v = e^{\mu}$. ■

E.5 Natural Gradient for the Multinomial Family: Full Derivation

The multinomial family on $\mathcal{X} = \{1, \dots, K\}$ has distributions $p = (p_1, \dots, p_K)$ in the open simplex Δ°_{K-1}. We work in the coordinates $\theta = (p_1, \dots, p_{K-1})$ (dropping $p_K = 1 - \sum_{j<K} p_j$). A loss function $L(p_1, \dots, p_{K-1})$ has ordinary gradient $\nabla L = (\partial L / \partial p_j)_{j<K}$.

The Fisher metric in p-coordinates:

$$g^F_{ij}(p) = \frac{\delta_{ij}}{p_i} + \frac{1}{p_K}, \quad i, j \in \{1, \dots, K-1\}.$$

The inverse Fisher metric (computed by block inversion or direct formula):

$$(g^F)^{ij}(p) = p_i(\delta_{ij} - p_j), \quad i, j \in \{1, \dots, K-1\}.$$

(This can be verified: $(g^F)^{ij} g^F_{jk} = p_i(\delta_{ij} - p_j)(\delta_{jk}/p_j + 1/p_K)$.)

The natural gradient in p-coordinates:

$$\tilde{\nabla}_{p_i} L = \sum_j (g^F)^{ij} \frac{\partial L}{\partial p_j} = p_i \frac{\partial L}{\partial p_i} - p_i \sum_j p_j \frac{\partial L}{\partial p_j}.$$

The natural gradient descent update $p_i \leftarrow p_i - \varepsilon \, \tilde{\nabla}_{p_i} L$:

$$p_i^{(t+1)} = p_i^{(t)} - \varepsilon \, p_i^{(t)} \left(\frac{\partial L}{\partial p_i}\bigg|_t - \sum_j p_j^{(t)} \frac{\partial L}{\partial p_j}\bigg|_t \right).$$

For the cross-entropy loss $L = -\sum_i q_i \log p_i$ (target distribution q): $\partial L / \partial p_i = -q_i / p_i$, so

$$\tilde{\nabla}_{p_i} L = p_i \cdot \left(-\frac{q_i}{p_i} \right) - p_i \sum_j p_j \cdot \left(-\frac{q_j}{p_j} \right) = -q_i + p_i \sum_j q_j = p_i - q_i.$$

The natural gradient update becomes $p^{(t+1)} = p^{(t)} - \varepsilon(p^{(t)} - q) = (1 - \varepsilon)p^{(t)} + \varepsilon q$: a linear interpolation (arithmetic mixture) between $p^{(t)}$ and the target q. This is precisely the m-geodesic from $p^{(t)}$ to q, confirming that natural gradient descent for cross-entropy loss on the multinomial simplex follows the mixture connection geodesic.

Closing Synthesis: The Geometry Behind Information Theory

The three volumes of this series chart a progression from the combinatorial and probabilistic foundations of information theory to its algebraic-algorithmic machinery, and finally to the differential-geometric structure that unifies both.

Volume I established Shannon's fundamental quantities — entropy, mutual information, channel capacity — as limits of combinatorial counting: the AEP shows that typical sets have size approximately 2^{nH}, and the channel coding theorem shows that 2^{nI} codewords can be reliably distinguished through n channel uses. These are asymptotic results, and the underlying distributions appear as *points* to be counted, not as geometric objects.

Volume II pushed inside Shannon's existence theorems: it asked how codes are constructed, how they are decoded, and what algebraic structures make them efficient. Reed-Solomon codes exploit the geometry of finite fields; LDPC codes exploit the sparsity of Tanner graphs; turbo codes exploit the product structure of component codes. The convergence of belief propagation in these settings is empirically excellent but theoretically subtle — and it is precisely here that information geometry provides the right language.

Volume III has developed the differential-geometric perspective. The key insights:

1. **Distributions are points on a manifold.** The Fisher metric is the unique Riemannian metric on statistical manifolds compatible with the data-processing inequality (Chentsov's theorem). It measures the distinguishability of nearby distributions.

2. **The manifold has two flat connection structures.** On exponential families, the e-connection (flat in θ) and m-connection (flat in η) coexist, linked by the Legendre duality of the log-partition function. This dual flatness is responsible for the exact Pythagorean theorem for KL divergence.

3. **Inference algorithms are geometric.** Maximum likelihood estimation is an e-projection (moment-matching), the EM algorithm alternates e- and m-projections, IPF alternates m-projections, and Bayesian updating is a mixture of prior and data in the m-flat coordinate system. All these algorithms are guaranteed to converge precisely because they are projections in a flat geometry.

4. **Optimisation on statistical manifolds requires the natural gradient.** The ordinary gradient is parameterisation-dependent; the natural gradient is not. K-FAC and related methods make the natural gradient computationally tractable for deep networks, and the Fisher-Rao gradient flow reduces to a linear ODE in the moment coordinates of exponential families.

Looking forward: the interplay between the three volumes is not exhausted by what has been covered. The information-geometric reinterpretation of LDPC belief propagation (as message passing on a Bethe free energy surface, which is a projection problem on a constrained manifold) and the geometry of turbo decoding connect Volumes II and III in ways that are still being actively explored. The α-family of divergences provides a one-parameter interpolation between the KL divergence (the channel coding bound of Volume I) and the squared Hellinger distance (related to the error exponent in hypothesis testing). The quantum extensions of all three volumes — quantum Shannon theory, quantum error correction, and quantum information geometry — are a major active research direction.

The reader who has completed all three volumes possesses the mathematical foundation to engage with this research frontier. The tools are in place: information theory provides the fundamental limits, coding theory provides the constructions, and information geometry

provides the geometric language in which both can be expressed, analysed, and extended.

F f-Divergences and Rényi Divergences

F.1 The f-Divergence Family

The KL divergence is not the only natural measure of dissimilarity between probability distributions. The *f-divergence* family (Ali and Silvey 1966; Csiszár 1967) provides a unified framework:

Definition F.1 — f-Divergence

Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. The *f-divergence* between distributions p and q (with $q \gg 0$) is

$$D_f(p\|q) = \int q(x)\, f\!\left(\frac{p(x)}{q(x)}\right) d\mu(x).$$

Key properties (all following from convexity of f and Jensen's inequality):

(i) $D_f(p\|q) \geq f(1) = 0$ (non-negativity), with equality iff $p = q$ (under strict convexity).

(ii) $D_f(\kappa_* p\|\kappa_* q) \leq D_f(p\|q)$ for any Markov kernel κ (data-processing inequality, i.e., processing cannot increase divergence).

(iii) The second-order Taylor expansion around $p = q$ gives $D_f(p\|q) \approx \frac{f''(1)}{2}\,\chi^2(p\|q)$, so all f-divergences have the same local geometry determined by the Fisher metric (up to a scalar multiple $f''(1)/2$).

Principal examples.

Name	$f(t)$	$D_f(p\|q)$	$f''(1)$
KL divergence	$t \log t$	$\int p \log(p/q)$	1
Reverse KL	$-\log t$	$\int q \log(q/p)$	1
χ^2-divergence	$(t-1)^2$	$\int (p-q)^2/q$	2
Squared Hellinger	$(\sqrt{t}-1)^2$	$1 - \int \sqrt{pq}$	1/2
Total variation	$\frac{1}{2}\lvert t-1\rvert$	$\frac{1}{2}\int \lvert p-q\rvert$	0 (non-smooth)
α-divergence	$\frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}$	$D^{(\alpha)}(p\|q)$	1

The *α-divergence* of Amari is the f-divergence with $f(t) = \frac{4}{1-\alpha^2}(1 - t^{(1+\alpha)/2})$ (alternative normalisation), giving:

$$D^{(\alpha)}(p\|q) = \frac{4}{1-\alpha^2}\left(1 - \int p^{(1+\alpha)/2} q^{(1-\alpha)/2}\, d\mu\right).$$

For $\alpha \to 1$: $D^{(1)}(p\|q) = D_{\mathrm{KL}}(p\,\|\,q)$. For $\alpha \to -1$: $D^{(-1)}(p\|q) = D_{\mathrm{KL}}(q\,\|\,p)$. For $\alpha = 0$: $D^{(0)}(p\|q) = 4(1 - \int \sqrt{pq}) = 2H^2(p, q)$ (twice squared Hellinger).

F.2 Rényi Divergences and Their Connection to α-Connections

Definition F.2 — Rényi Divergence

For $\alpha \in (0,1) \cup (1,\infty)$, the *Rényi divergence of order* α is

$$D_\alpha(p\|q) = \frac{1}{\alpha-1} \log \int p(x)^\alpha q(x)^{1-\alpha}\, d\mu(x) = \frac{1}{\alpha-1} \log \mathbb{E}_q\left[\left(\frac{p}{q}\right)^\alpha\right].$$

Key relations:

- $\lim_{\alpha\to 1} D_\alpha(p\|q) = D_{\mathrm{KL}}(p\,\|\,q)$ (by L'Hôpital).
- $\lim_{\alpha\to 0} D_\alpha(p\|q) = -\log \int \mathbf{1}[p>0]\, q\, d\mu$ (Rényi entropy of order 0).
- D_α is non-decreasing in α: $D_\alpha \le D_\beta$ for $\alpha \le \beta$.
- Data-processing inequality: $D_\alpha(\kappa_* p\|\kappa_* q) \le D_\alpha(p\|q)$.

Connection to hypothesis testing. In the Neyman-Pearson problem of testing $H_0 : p = p_0$ vs. $H_1 : p = p_1$ at error level ε, the optimal type-II error exponent is given by Stein's lemma: $-\frac{1}{n}\log \beta_n \to D_{\mathrm{KL}}(p_1\,\|\,p_0)$. For the two-sided problem, the error exponent is the *Chernoff information*:

$$C(p_0,p_1) = -\min_{\alpha\in(0,1)} \log \int p_0^\alpha p_1^{1-\alpha}\, d\mu = \max_{\alpha\in(0,1)} ((\alpha-1)D_\alpha(p_1\|p_0)).$$

The Chernoff information is therefore the maximum Rényi divergence (suitably normalised) over the unit interval, and it governs the optimal simultaneous control of type-I and type-II errors.

Information-geometric interpretation. The family of distributions $\{p_0^\alpha p_1^{1-\alpha}/Z(\alpha)\}$ for $\alpha \in [0,1]$ is an *e-geodesic* from p_1 (at $\alpha = 0$) to p_0 (at $\alpha = 1$) in the exponential connection — it is a one-parameter exponential family in the natural parameter $\alpha \log(p_0/p_1)$. The Rényi divergence D_α is the log-partition function of this family. The Chernoff information is the maximum of the Rényi divergence and is achieved at the midpoint of the e-geodesic that is equidistant (in KL divergence) from both endpoints.

F.3 Second-Order Geometry and the Metric from f-Divergences

All f-divergences with $f''(1) > 0$ induce the same Riemannian metric (up to a scalar) on any statistical manifold. This is because for p_θ and $p_{\theta+d\theta}$:

$$\begin{aligned} D_f(p_{\theta+d\theta}\|p_\theta) &= \int p_\theta(x) f\left(\frac{p_{\theta+d\theta}(x)}{p_\theta(x)}\right) d\mu \\ &\approx \int p_\theta \left[f(1) + f'(1)\left(\frac{p_{\theta+d\theta}-p_\theta}{p_\theta}\right) + \frac{f''(1)}{2}\left(\frac{p_{\theta+d\theta}-p_\theta}{p_\theta}\right)^2 \right] d\mu \\ &= 0 + f'(1)\underbrace{\int (p_{\theta+d\theta}-p_\theta) d\mu}_{=0} + \frac{f''(1)}{2}\int \frac{(p_{\theta+d\theta}-p_\theta)^2}{p_\theta}\, d\mu. \end{aligned}$$

Now $\int (p_{\theta+d\theta}-p_\theta)^2/p_\theta\, d\mu = \sum_{i,j} g^F_{ij} d\theta^i d\theta^j + O(|d\theta|^3)$, so:

$$D_f(p_{\theta+d\theta}\|p_\theta) = \frac{f''(1)}{2}\sum_{i,j} g^F_{ij} d\theta^i d\theta^j + O(|d\theta|^3).$$

The second-order structure of any f-divergence is $\frac{f''(1)}{2}$ times the Fisher metric. This is a rigorous statement of Chentsov's uniqueness theorem at the infinitesimal level: the Fisher metric is the *only* metric that arises from the second-order expansion of any f-divergence.

F.4 The Geometry of Rényi Entropy

Rényi entropy of order α for a distribution p on a finite alphabet $\mathcal{X}$:

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} p(x)^\alpha.$$

Limits: $H_0(p) = \log |\mathrm{supp}(p)|$ (Hartley entropy), $H_1(p) = H(p)$ (Shannon entropy, by L'Hôpital), $H_\infty(p) = -\log \max_x p(x)$ (min-entropy). Rényi entropy is non-increasing in α: $H_\alpha \geq H_\beta$ for $\alpha \leq \beta$.

The Rényi entropy appears in information-theoretic security (min-entropy governs extractable randomness), in physics (Tsallis entropy is a reparametrisation of Rényi entropy), and in the analysis of quantum channels (quantum Rényi divergences and Rényi channel capacity).

The *escort distribution* associated with p at order α is $p^{[\alpha]}(x) = p(x)^\alpha / Z_\alpha$, where $Z_\alpha = \sum_x p(x)^\alpha$. The map $p \mapsto p^{[\alpha]}$ is an e-connection geodesic path: the family $\{p^{[\alpha]} : \alpha \in \mathbb{R}\}$ for fixed p is a one-parameter exponential family with sufficient statistic $T(x) = \log p(x)$ and natural parameter α.

Tsallis entropy and q-deformed statistics. The *Tsallis entropy* of order $q > 0$, $q \neq 1$, is:

$$S_q(p) = \frac{1}{q-1}\left(1 - \sum_x p(x)^q\right) = \frac{1 - \sum_x p_x^q}{q-1}.$$

It reduces to Shannon entropy as $q \to 1$: $\lim_{q\to 1} S_q(p) = H(p)$. The connection to Rényi entropy is: $S_q(p) = (1 - e^{(1-q)H_q(p)})/(q-1)$ (a monotone function of H_q for fixed q). Tsallis entropy is the objective in *q-deformed maximum entropy* (non-extensive statistical mechanics, Tsallis 1988): instead of $\max H(p)$ subject to $\mathbb{E}[f] = \mu$, one solves $\max S_q(p)$ subject to $\mathbb{E}_{p^{[q]}}[f] = \mu$ (the constraint is w.r.t. the escort distribution).

The solution is the *q-exponential distribution*:

$$p_q^*(x) \propto [1 + (q-1)\lambda^\top f(x)]_+^{1/(1-q)},$$

where $[u]_+ = \max(u, 0)$. For $q \to 1$: $p_q^* \to \exp(\lambda^\top f)$ (the standard Boltzmann/exponential family distribution). For $q > 1$: p_q^* has power-law tails $\propto x^{-1/(q-1)}$ (heavy tails); for $q < 1$: p_q^* has compact support (tails truncated to zero). The family $\{p_q^*(\cdot;\lambda) : q, \lambda\}$ is an *α-exponential family* in the sense of Amari-Nagaoka ($\alpha = 1 - 2q$), and its geometry is governed by the α-connections of Part 3 at the corresponding value of α.

Geometric programming and log-convex models. The class of *log-linear models* $p_\theta(x) \propto \exp(\theta^\top f(x))$ (standard exponential families) can be extended to *log-convex models* $p(x) \propto g(\theta, f(x))$ where $\log g$ is convex in θ, which includes the q-exponential family when $q \neq 1$. The maximum likelihood problem for such models is a *geometric program* (a convex optimisation problem after taking logarithms), solvable in polynomial time. The information-geometric structure for these "generalised exponential families" parallels the standard theory: there is a generalised log-partition function $\Phi_q(\theta) = \log \sum_x [1+(q-1)\theta^\top f(x)]^{q/(q-1)}$ (the q-deformed log-partition), its gradient gives generalised moments $\mathbb{E}_{p^{[q]}}[f]$, and its Hessian gives a generalised Fisher metric. The resulting geometry is a dually-flat Hessian metric, but in a different coordinate system from the standard (θ, η) duality of Parts 2–3.

The Rényi divergence $D_\alpha(p\|q)$ for $\alpha \in (0,1)$ can be written as:

$$D_\alpha(p\|q) = \frac{1}{\alpha-1}\log \mathbb{E}_q\left[\left(\frac{p}{q}\right)^\alpha\right] = \frac{1}{\alpha-1}\log \int p(x)^\alpha q(x)^{1-\alpha}\, d\mu(x).$$

The integrand $\int p^\alpha q^{1-\alpha} d\mu$ is the *Bhattacharyya coefficient* raised to the power α (for $\alpha = 1/2$: $\int \sqrt{pq}\, d\mu = \mathrm{BC}(p,q)$, and $D_{1/2}(p\|q) = -2\log \mathrm{BC}(p,q) = 2H^2(p,q)/2$, related to the squared Hellinger distance $H^2(p,q) = \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2 d\mu$). The family $\{p^\alpha q^{1-\alpha}/Z(\alpha)\}_{\alpha\in[0,1]}$ is the e-geodesic from q to p, and $\log Z(\alpha) = -(\alpha-1)D_\alpha(p\|q)$ is the log-partition function of this path. Thus D_α is the cumulant generating function (at "time" α) of the log-likelihood ratio $\log(p/q)$ under the exponential tilting $p^\alpha q^{1-\alpha}/Z(\alpha)$.

G Connections to Volume II: Belief Propagation as Geometric Projection

G.1 The Bethe Free Energy and m-Projections

In Volume II, belief propagation (the sum-product algorithm) on a Tanner graph was presented as a message-passing procedure for computing marginals of a factor graph. Information geometry provides a deeper analysis: belief propagation is a coordinate ascent on the *Bethe free energy*, and the fixed points of belief propagation are m-projections onto the Bethe approximation manifold.

Let $G = (\mathcal{V}, \mathcal{F}, \mathcal{E})$ be a factor graph with variable nodes $i \in \mathcal{V}$, factor nodes $a \in \mathcal{F}$, and edges $\mathcal{E}$. A joint distribution factorises as $p(x) = \prod_a f_a(x_a)/Z$, where x_a denotes the variables adjacent to factor a. The *Bethe free energy* is

$$F_{\text{Bethe}}(\{b_a\}, \{b_i\}) = \sum_a \sum_{x_a} b_a(x_a) \log \frac{b_a(x_a)}{f_a(x_a)} - \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \log b_i(x_i),$$

where $\{b_a\}$ and $\{b_i\}$ are local beliefs on factors and variables respectively, constrained to be consistent: $\sum_{x_a \setminus i} b_a(x_a) = b_i(x_i)$ for all $i \in \partial a$. Here d_i is the degree of variable i.

The Bethe free energy approximates the true free energy $F = D_{\mathrm{KL}}\big(p^{\text{model}} \,\|\, p^{\text{true}}\big) + \text{const}$ by replacing the global entropy with a sum of local entropies: $H(\{b_a\}, \{b_i\}) \approx \sum_a H(b_a) - \sum_i (d_i - 1) H(b_i)$, which is exact on trees (where the Bethe approximation is exact) and an approximation for loopy graphs.

Information-geometric formulation. The constraint set $\mathcal{L} = \{\{b_a, b_i\} : \text{consistency}\}$ is an m-flat submanifold of the product simplex $\prod_a \mathcal{P}(\mathcal{X}_a) \times \prod_i \mathcal{P}(\mathcal{X}_i)$. Minimising F_{Bethe} over $\mathcal{L}$ is an m-projection of the model p onto $\mathcal{L}$:

$$\{b_a^*, b_i^*\} = \pi^{(m)}(p \to \mathcal{L}) = \operatorname*{arg\,min}_{\{b\} \in \mathcal{L}} D_{\mathrm{KL}}(\{b\} \,\|\, p)\,.$$

The belief propagation messages $\{m_{i \to a}, m_{a \to i}\}$ are the Lagrange multipliers for the consistency constraints, and their fixed-point equations are precisely the stationary conditions for the m-projection.

G.2 Convergence and the Geometry of Loopy BP

On a tree (no cycles), the Bethe approximation is exact ($\mathcal{L}$ intersects the model manifold at a single point), and belief propagation converges in a number of iterations equal to the diameter of the tree. The convergence is geometric with rate 0 (finite termination), corresponding to the fact that the two submanifolds being projected between are orthogonal at the solution.

On a loopy graph (cycles), the Bethe manifold $\mathcal{L}$ may not intersect the model manifold, and belief propagation may not converge. The non-convergence is a geometric phenomenon: the alternating projections are not contractions in the Fisher metric because the two submanifolds are not flat in compatible coordinate systems. Several conditions for convergence are known:

(i) **Walk-summability** (Malioutov et al. 2006): BP converges for Gaussian graphical models with spectral radius $\rho(|J| - \mathrm{diag}(|J|)) < 1$, where J is the information matrix. This is a curvature condition on the model manifold.

(ii) **Contractivity in the Hilbert metric** (Mooij and Kappen 2007): BP messages converge if the factor nodes define contractive maps on the positive cone with respect to the Hilbert projective metric, which is a non-symmetric version of the Fisher metric.

(iii) **Decay of correlations** (Weitz 2006, for discrete models): BP converges on graphs of bounded degree when the model is in the "unique Gibbs measure" phase (below the phase transition threshold), which corresponds geometrically to the model manifold being sufficiently "flat" near the target distribution.

G.3 LDPC Density Evolution as a Fixed-Point Equation on a Manifold

In Volume II, density evolution for LDPC codes described the propagation of error probability distributions through the belief propagation decoder. Information-geometrically, density evolution is a fixed-point equation on the space of probability distributions on $\mathbb{R}$ (the log-likelihood ratios): at each iteration, the distribution of LLRs transforms as a convolution (check-node operation, corresponding to an m-mixture in LLR space) followed by a nonlinear transform (variable-node operation).

The capacity-achieving condition (density evolution threshold) is the condition under which the fixed-point distribution is the delta mass at $+\infty$ (perfect decoding). This is a bifurcation in the dynamical system on the manifold of LLR distributions, and it occurs precisely when the manifold of achievable LLR distributions at the decoder output intersects the "good" region of the simplex.

The geometric picture unifies the algebraic (finite-field) structure of LDPC codes (from Volume II) with the probabilistic geometry of information theory (Volume I) and the differential-geometric framework of this volume: the code is a linear subspace (an affine e-flat submanifold) of the binary product space, and iterative decoding alternates projections onto this subspace and onto the "received signal manifold" determined by the channel transition probabilities.

H Statistical Manifolds in Machine Learning: Extended Examples

H.1 Variational Inference as m-Projection

Variational inference (VI) is a method for approximating a posterior distribution $p(\theta|x) \propto p(x|\theta)p(\theta)$ by a distribution q_ϕ in a tractable family $\mathcal{Q} = \{q_\phi : \phi \in \Phi\}$ (e.g., a mean-field Gaussian family). The VI objective is to minimise the KL divergence:

$$\phi^* = \arg\min_{\phi \in \Phi} D_{\mathrm{KL}}\big(q_\phi \,\|\, p(\cdot|x)\big)\,.$$

This is precisely the m-*projection* of the true posterior $p(\cdot|x)$ onto the variational family $\mathcal{Q}$.

The ELBO as a KL Decomposition

The evidence lower bound (ELBO) is:

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\log p(x,\theta)] - \mathbb{E}_{q_\phi}[\log q_\phi(\theta)] = \log p(x) - D_{\mathrm{KL}}\big(q_\phi \,\|\, p(\cdot|x)\big)\,.$$

Maximising the ELBO is equivalent to minimising $D_{\mathrm{KL}}\big(q_\phi \,\|\, p(\cdot|x)\big)$, confirming the m-projection interpretation. Since $\log p(x)$ is constant in ϕ, $\max_\phi \mathcal{L}(\phi) = \log p(x) - \min_\phi D_{\mathrm{KL}}\big(q_\phi \,\|\, p(\cdot|x)\big)$.

Why m-projection (not e-projection)? The choice of $D_{\mathrm{KL}}(q \,\|\, p)$ rather than $D_{\mathrm{KL}}(p \,\|\, q)$ has geometric consequences. The m-projection (minimising $D_{\mathrm{KL}}(q \,\|\, p)$) is "zero-forcing": it makes q small wherever p is small, leading to approximations that are unimodal when the posterior p is multimodal. The e-projection (minimising $D_{\mathrm{KL}}(p \,\|\, q)$, i.e., expectation propagation) is "mass-covering": it seeks q that covers all modes of p, leading to broader approximations that do not "collapse" to a single mode.

Mean-field factorisation as alternating m-projections. The mean-field family $\mathcal{Q}_{\mathrm{MF}} = \{q : q(\theta) = \prod_i q_i(\theta_i)\}$ is an m-flat submanifold of $\mathcal{P}(\Theta)$ (it is the closure of the family of product distributions under m-mixtures). The coordinate ascent VI algorithm alternates m-projections onto the submanifolds $\{q : q_{j\neq i}$ fixed$\}$ for each factor:

$$q_i^*(\theta_i) \propto \exp\Big(\mathbb{E}_{q_{-i}}[\log p(x,\theta)]\Big)\,.$$

This is the m-projection of $p(\cdot|x)$ onto the mean-field submanifold with all other factors fixed, and coordinate ascent VI converges by the same Csiszár–Tusnády argument as the EM algorithm.

H.3 Normalising Flows and the Geometry of Transport Maps

A *normalising flow* defines a distribution p_ϕ on $\mathbb{R}^d$ as the pushforward of a simple base distribution p_0 (e.g., $\mathcal{N}(0, I)$) through an invertible parametric map $f_\phi : \mathbb{R}^d \to \mathbb{R}^d$:

$$p_\phi(x) = p_0(f_\phi^{-1}(x))\,|\det J_{f_\phi^{-1}}(x)|,$$

where $J_{f_\phi^{-1}}$ is the Jacobian. The model family $\{p_\phi : \phi \in \Phi\}$ is in general *neither* an exponential family *nor* a mixture family: it is a curved submanifold of $\mathcal{P}(\mathbb{R}^d)$ that is neither e-flat nor m-flat.

Information-geometric structure. The Fisher information of the flow model is

$$\mathbf{G}_{ij}^{(\phi)} = \mathbb{E}_{p_\phi}\big[\partial_i \log p_\phi(X)\, \partial_j \log p_\phi(X)\big]\,,$$

which involves second-order derivatives of $\log\det J_{f_\phi}$ (via the log-density formula). For autoregressive flows (where f_ϕ is triangular and $\det J_{f_\phi} = \prod_k \partial_k f_k$), the Fisher information can be decomposed layer-by-layer, but it does not factorise cleanly: the curvature tensor of $\{p_\phi\}$ in $\mathcal{P}(\mathbb{R}^d)$ is generally nonzero in all three connections (e, m, and Levi-Civita), making normalising flows a class of *curved* statistical manifolds for which the dually-flat machinery does not directly apply.

Training as m-projection. Maximum likelihood training of a normalising flow minimises $D_{\mathrm{KL}}\big(\hat{P}_n \,\|\, p_\phi\big)$ (the e-projection of the empirical distribution $\hat{P}_n$ onto $\{p_\phi\}$), exactly as for exponential families. The difference is that the M-step of the gradient ascent is not analytically available (the flow lacks closed-form moment-matching), and one must use stochastic gradient ascent on the log-likelihood:

$$\nabla_\phi \log p_\phi(x) = \nabla_\phi \log p_0(f_\phi^{-1}(x)) + \nabla_\phi \log|\det J_{f_\phi^{-1}}(x)|.$$

The natural gradient for the flow is $\mathbf{G}^{(\phi)-1}\nabla_\phi \log p_\phi$, but computing $\mathbf{G}^{(\phi)}$ and its inverse is as expensive as for general neural networks: $O(d^2)$ to form and $O(d^3)$ to invert, making exact natural gradient impractical and requiring approximations (K-FAC, diagonal Fisher, or Riemannian adaptive methods).

H.4 Score Matching and Stein Operators on the Statistical Manifold

Score matching (Hyvärinen 2005) provides an alternative to maximum likelihood for distributions that are known up to a normalising constant: $p_\theta(x) = \tilde{p}_\theta(x)/Z(\theta)$, where $Z(\theta)$ is intractable.

The *score matching objective* is:

$$J_{\mathrm{SM}}(\theta) = \frac{1}{2}\mathbb{E}_{p^*}[\|\nabla_x \log p_\theta(x) - \nabla_x \log p^*(x)\|^2] = \mathbb{E}_{p^*}\left[\sum_k \partial_{x_k}^2 \log \tilde{p}_\theta + \tfrac{1}{2}\|\nabla_x \log \tilde{p}_\theta\|^2\right] + \text{const},$$

where the second form (after integration by parts) does not require $\nabla_x \log p^*(x)$ (the true score, which is intractable for an unnormalised model). Minimising $J_{\mathrm{SM}}(\theta)$ over θ provides an estimator that is consistent and asymptotically normal, with Fisher information $\mathbf{G}_{\mathrm{SM}}(\theta^*)$ replacing the standard $\mathbf{G}(\theta^*)$.

Information-geometric interpretation. The score matching objective can be written as:

$$J_{\mathrm{SM}}(\theta) = \tfrac{1}{2}\|\nabla_x \log p_\theta - \nabla_x \log p^*\|^2_{L^2(p^*)} = \tfrac{1}{2}\|s_\theta - s^*\|^2_{L^2(p^*)},$$

a squared $L^2(p^*)$-distance between the *scores* $s_\theta = \nabla_x \log p_\theta$ and $s^* = \nabla_x \log p^*$. The space of score functions $\{s_\theta : \theta \in \Theta\}$ is a submanifold of the (infinite-dimensional) Hilbert space $L^2(p^*)$, and score matching is the projection of s^* onto this submanifold in the $L^2(p^*)$ metric. The connection to the Fisher metric is via the identity:

$$\mathbf{G}_{ij}(\theta) = \langle \partial_i \log p_\theta,\, \partial_j \log p_\theta \rangle_{L^2(p_\theta)} = \mathbb{E}_{p_\theta}[\partial_i \log p_\theta \cdot \partial_j \log p_\theta],$$

so the Fisher metric on the parameter manifold is the Gram matrix of the score functions in the $L^2(p_\theta)$ inner product. Score matching uses a different (fixed) inner product $L^2(p^*)$, giving a different (but related) metric on the parameter space.

H.5 Diffusion Models and the Wasserstein-Fisher-Rao Geometry

Denoising diffusion probabilistic models (DDPMs, Ho et al. 2020) define a forward process $q(x_t|x_0) = \mathcal{N}(x_0\sqrt{\bar{\alpha}_t}, (1-\bar{\alpha}_t)I)$ that gradually adds Gaussian noise, and a reverse process

$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$ that learns to denoise. The training objective is a variational bound on $D_{\mathrm{KL}}(q(x_{0:T}) \,\|\, p_\theta(x_{0:T}))$:

$$\mathcal{L}_{\mathrm{DDPM}}(\theta) = \mathbb{E}_q\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(q(x_{t-1}|x_t,x_0) \,\|\, p_\theta(x_{t-1}|x_t))\right].$$

Each term $D_{\mathrm{KL}}(q(\cdot|x_t,x_0) \,\|\, p_\theta(\cdot|x_t))$ is a KL divergence between two Gaussians and is equal to the Fisher-Rao squared distance (since for Gaussians, KL divergence is a function of the Fisher-Rao metric on $\mathcal{N}(\mu,\sigma^2 I)$). DDPM training is thus a sum of m-projection steps, one per diffusion time step, projecting the analytical reverse posterior $q(\cdot|x_t,x_0)$ onto the parametric Gaussian model $p_\theta(\cdot|x_t)$.

The resulting squared loss on the *noise prediction* (ϵ-parametrisation):

$$\mathcal{L}_{\mathrm{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t,t)\|^2]$$

is a Euclidean squared loss on the *score function* (since $\epsilon_\theta(x_t,t) \approx -\sqrt{1-\bar{\alpha}_t}\nabla_{x_t}\log p_\theta(x_t)$), connecting DDPM training to score matching: diffusion model training is a time-conditioned score matching objective, evaluated at different noise levels.

> The Wasserstein-Fisher-Rao (WFR) metric (also called the Hellinger-Kantorovich metric, Chizat et al. 2018) interpolates between the Fisher-Rao metric (Section 1.4) and the 2-Wasserstein metric of optimal transport, unifying two complementary geometric frameworks: Fisher-Rao measures "statistical distinguishability" (how easily two distributions can be discriminated by a test), while Wasserstein measures "transport cost" (how much mass must be moved to transform one distribution into the other). The WFR geodesic between two distributions is neither the e-geodesic nor the m-geodesic of information geometry but a curve in the space of measures that simultaneously changes mass (like Wasserstein) and changes the shape of the density (like Fisher-Rao). The WFR metric and its geodesic equations are the subject of active research, with applications to generative modelling, optimal control of probability distributions, and the geometry of population dynamics. Diffusion models and normalising flows can be seen as approximate WFR geodesic flows, with the diffusion noise schedule governing the interpolation between the Wasserstein and Fisher-Rao regimes.

The *variational autoencoder* (VAE, Kingma and Welling 2013) performs variational inference with an amortised encoder: instead of optimising ϕ for each datapoint x, it learns a function $\phi(x)$ (the encoder network) that maps each x to the variational parameters. The VAE objective for a single datapoint x:

$$\mathcal{L}(\phi,\psi;x) = \mathbb{E}_{q_{\phi(x)}(z)}[\log p_\psi(x|z)] - D_{\mathrm{KL}}\big(q_{\phi(x)}(z) \,\|\, p(z)\big),$$

where $q_{\phi(x)}(z) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$ (encoder) and $p_\psi(x|z)$ (decoder) is parametrised by ψ.

Information-geometrically, the VAE performs a *simultaneous* m-projection (in ϕ) and maximum likelihood estimation (in ψ), coupled through the reparametrisation trick. The amortisation means the projection manifold is the image of the encoder: the set of Gaussians $\{\mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) : x \in \mathcal{X}\}$, which is a *learned* approximating manifold rather than a fixed variational family.

The natural gradient for the encoder parameters $\phi(x)$ (at a given x) is computed using the Fisher information of $q_{\phi(x)}$:

$$\tilde{\nabla}_\phi \mathcal{L} = I(q_{\phi(x)})^{-1}\nabla_\phi \mathcal{L}.$$

For the Gaussian encoder $q = \mathcal{N}(\mu, \Sigma)$ with $\phi = (\mu, \Sigma)$, the Fisher metric is block-diagonal: $I = \mathrm{diag}(\Sigma^{-1}, \frac{1}{2}\Sigma^{-1} \otimes \Sigma^{-1})$. The natural gradient descent for μ becomes $\mu \leftarrow \mu - \varepsilon \Sigma \nabla_\mu \mathcal{L}$, which is the Newton step for the mean of a Gaussian encoder — scaling the gradient by the covariance, not by a scalar learning rate.

H.3 Normalising Flows and the Geometry of Change of Variables

A *normalising flow* defines a probability distribution p_θ on $\mathbb{R}^d$ by transforming a base distribution p_0 (e.g., $\mathcal{N}(0, I)$) through a learnable diffeomorphism $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$:

$$p_\theta(x) = p_0(f_\theta^{-1}(x)) \, |\det J f_\theta^{-1}(x)|.$$

The log-likelihood is $\log p_\theta(x) = \log p_0(f_\theta^{-1}(x)) + \log |\det J f_\theta^{-1}(x)|$.

Information geometry of flows. The statistical manifold $\mathcal{M}_\theta = \{p_\theta : \theta \in \Theta\}$ is a submanifold of the space $\mathcal{P}(\mathbb{R}^d)$ of all probability distributions. The Fisher information matrix for θ is:

$$\begin{aligned} I(\theta)_{ij} &= \mathbb{E}_{p_\theta}\left[\frac{\partial \log p_\theta}{\partial \theta^i} \frac{\partial \log p_\theta}{\partial \theta^j}\right] \\ &= \mathbb{E}_{z \sim p_0}\left[\frac{\partial}{\partial \theta^i} \log p_\theta(f_\theta(z)) \cdot \frac{\partial}{\partial \theta^j} \log p_\theta(f_\theta(z))\right], \end{aligned}$$

where the last step uses the change-of-variables formula and $z = f_\theta^{-1}(x)$.

The natural gradient for normalising flow training (maximum likelihood on a dataset $\{x_k\}$):

$$\theta_{t+1} = \theta_t + \varepsilon \, I(\theta_t)^{-1} \, \frac{1}{n} \sum_{k=1}^{n} \nabla_\theta \log p_{\theta_t}(x_k),$$

is the parametrisation-invariant steepest ascent on the log-likelihood surface, generalising the natural gradient for exponential families to the non-exponential flow manifold. While the Fisher matrix is dense and expensive to invert for flow models, diagonal or Kronecker approximations (analogous to K-FAC) can be applied to the Jacobian terms.

H.4 Score Matching and the Fisher Divergence

Score matching (Hyvärinen 2005) is an alternative to maximum likelihood for training energy-based models $p_\theta(x) \propto \exp(-E_\theta(x))$ when the partition function $Z(\theta) = \int \exp(-E_\theta)$ is intractable. It minimises the *Fisher divergence*:

$$J(\theta) = \frac{1}{2} \mathbb{E}_{p_{\text{data}}}\left[\|\nabla_x \log p_{\text{data}}(x) - \nabla_x \log p_\theta(x)\|^2\right].$$

By integration by parts (Hyvärinen's trick):

$$J(\theta) = \mathbb{E}_{p_{\text{data}}}\left[\sum_i \partial_{x_i}^2 \log p_\theta(x) + \frac{1}{2} \|\nabla_x \log p_\theta(x)\|^2\right] + \text{const},$$

eliminating the score of p_{data} (which is unknown) and the intractable partition function.

Information-geometric interpretation. The Fisher divergence $J(\theta)$ is a discrepancy in the *score functions* (tangent vectors on the data manifold), rather than in the distributions themselves. It is related to the Fisher information metric in the following sense: the Fisher information matrix $I(\theta) = \mathbb{E}_{p_\theta}[\nabla_\theta \log p_\theta \cdot (\nabla_\theta \log p_\theta)^\top]$ measures the sensitivity of the distribution

to parameter changes, while the Fisher divergence measures the sensitivity of the score function to x (the Hessian-type term $\nabla_x^2 \log p_\theta$).

Score matching is the basis of *denoising diffusion models* (Ho et al. 2020, Song et al. 2020), which learn the score function $\nabla_x \log p(x|t)$ (the gradient of the log-density of the noised data at noise level t) and use it to run a reverse diffusion process. Information-geometrically, the reverse diffusion is a time-reversal of the Ornstein-Uhlenbeck process, which is a gradient flow on the Wasserstein manifold — the L^2-Riemannian structure on $\mathcal{P}(\mathbb{R}^d)$ complementary to the Fisher-Rao structure.

H.5 Geometry of the Softmax Temperature and Entropy Regularisation

In reinforcement learning and language models, the softmax output is often parametrised with a temperature $\tau > 0$:

$$p_\tau(a) = \frac{\exp(f(a)/\tau)}{\sum_{a'} \exp(f(a')/\tau)}.$$

As $\tau \to 0^+$: $p_\tau \to$ argmax (greedy policy). As $\tau \to \infty$: $p_\tau \to$ Uniform.

Temperature as e-geodesic parameter. The family $\{p_\tau : \tau > 0\}$ is a one-parameter exponential family with natural parameter $1/\tau$ and sufficient statistic $T(a) = f(a)$:

$$\log p_\tau(a) = \frac{1}{\tau} f(a) - \log Z(\tau).$$

Setting $\theta = 1/\tau$, the family is an e-geodesic in the exponential coordinate through the direction $f = (f(a))_{a \in \mathcal{A}}$. The temperature τ parametrises the position along this geodesic: high temperature is near the uniform distribution (the maximum entropy point), and low temperature is near the greedy argmax.

Entropy regularisation and m-projections. Maximum-entropy reinforcement learning optimises:

$$J_{\text{MaxEnt}}(\pi) = \sum_t \mathbb{E}[R(s_t, a_t) + \tau H(\pi(\cdot|s_t))] \,.$$

The entropy regularisation $\tau H(\pi)$ can be interpreted as a Bregman divergence: $H(\pi) = -D_{\text{KL}}(\pi \,\|\, \text{Uniform}) + \text{const}$, so the MaxEnt objective is the m-projection of the reward-optimal policy onto a ball of radius determined by τ in KL divergence from the uniform policy. The optimal policy has the softmax form $\pi^*(a|s) \propto \exp(Q^*(s,a)/\tau)$, where Q^* is the soft Q-function — confirming that the optimal MaxEnt policy is an e-geodesic from the uniform at distance $\tau^{-1}\|Q^*\|$ in the exponential direction.

I Worked Problems with Full Solutions

Unlike exercise sections with stated problems only, this appendix provides fully worked problems with detailed solutions that extend the theoretical development of the main text. Each problem is chosen to illuminate a non-obvious aspect of the theory.

Problem I.1: Fisher Metric on the Poisson Family

Problem. Compute the Fisher information metric on the Poisson family $\{p_\lambda : \lambda > 0\}$ with $p_\lambda(k) = e^{-\lambda}\lambda^k/k!$ for $k \in \{0, 1, 2, ...\}$. Show that the Poisson family is a one-dimensional exponential family, identify its natural and moment parameters, and compute the geodesic distance between p_{λ_1} and p_{λ_2}.

Solution. The log-likelihood is $\log p_\lambda(k) = -\lambda + k\log\lambda - \log k!$. This can be written as $\log p_\lambda(k) = \theta k - e^\theta - \log k!$ where $\theta = \log\lambda$. This is an exponential family with natural parameter $\theta \in \mathbb{R}$ and sufficient statistic $T(k) = k$. The log-partition function is $\psi(\theta) = e^\theta = \lambda$.

Fisher information. Using the Bartlett identity:

$$-\frac{d^2}{d\lambda^2}\log p_\lambda(k) = \frac{k}{\lambda^2}.$$

Taking the expectation: $I(\lambda) = \mathbb{E}_\lambda[K/\lambda^2] = \lambda/\lambda^2 = 1/\lambda$.

In natural coordinates $\theta = \log\lambda$: $I(\theta) = \psi''(\theta) = e^\theta = \lambda$.

Moment parameter. $\eta = \psi'(\theta) = e^\theta = \lambda = \mathbb{E}_\lambda[K]$. So $\eta = \lambda$ (the mean), and $\psi^*(\eta) = \eta\log\eta - \eta$ (negative entropy of the Poisson).

Geodesic distance in the Fisher metric:

$$d_F(\lambda_1, \lambda_2) = \int_{\lambda_1}^{\lambda_2} \sqrt{I(\lambda)}\, d\lambda = \int_{\lambda_1}^{\lambda_2} \frac{1}{\sqrt{\lambda}}\, d\lambda = 2(\sqrt{\lambda_2} - \sqrt{\lambda_1}).$$

The Fisher-Rao geodesic in λ-coordinates is the curve $\gamma(t) = (\sqrt{\lambda_1}(1-t) + \sqrt{\lambda_2}\, t)^2$, an arc parametrised by the Bhattacharyya coefficient: at $t = 1/2$, $\gamma(1/2) = \left(\frac{\sqrt{\lambda_1}+\sqrt{\lambda_2}}{2}\right)^2$, which is not the arithmetic mean $(\lambda_1 + \lambda_2)/2$.

Verification via KL divergence.

$$D_{\mathrm{KL}}\big(p_{\lambda_1} \,\|\, p_{\lambda_2}\big) = \lambda_2 - \lambda_1 + \lambda_1\log(\lambda_1/\lambda_2) = B_{\psi^*}(\eta_2\|\eta_1) = \lambda_2 - \lambda_1 - \lambda_1\log(\lambda_2/\lambda_1).$$

The local expansion: $D_{\mathrm{KL}}(p_\lambda \,\|\, p_{\lambda+d\lambda}) \approx \frac{(d\lambda)^2}{2\lambda} = \frac{1}{2}I(\lambda)(d\lambda)^2$. Consistent with $I(\lambda) = 1/\lambda$. ■

Problem I.2: Pythagorean Theorem for Three Bernoulli Distributions

Problem. Let $p = \text{Bernoulli}(1/4)$, $q = \text{Bernoulli}(1/2)$, $r = \text{Bernoulli}(3/4)$. Show that the m-geodesic from p to r passes through q, that the generalised Pythagorean identity holds with an appropriate "hypotenuse", and verify numerically.

Solution. The Bernoulli family is a one-dimensional exponential family with natural parameter $\theta = \log(p/(1-p))$ (log-odds) and moment parameter $\eta = p$ (probability of 1). The m-geodesic from $p = 1/4$ to $r = 3/4$ in η-coordinates is the straight line $\eta(t) = (1-t)\cdot\frac{1}{4} + t\cdot\frac{3}{4}$. At $t = 1/2$: $\eta(1/2) = 1/2 = q$. So indeed q lies on the m-geodesic from p to r at its midpoint.

KL divergences.

$$D_{\mathrm{KL}}(p\,\|\,r) = \frac{1}{4}\log\frac{1/4}{3/4} + \frac{3}{4}\log\frac{3/4}{1/4} = \frac{1}{4}\log\frac{1}{3} + \frac{3}{4}\log 3 = \frac{1}{2}\log 3 \approx 0.549 \text{ nats.}$$

$$D_{\mathrm{KL}}(p\,\|\,q) = \frac{1}{4}\log\frac{1/4}{1/2} + \frac{3}{4}\log\frac{3/4}{1/2} = -\frac{1}{4}\log 2 + \frac{3}{4}\log\frac{3}{2} \approx 0.082 \text{ nats.}$$

$$D_{\mathrm{KL}}(q\,\|\,r) = \frac{1}{2}\log\frac{1/2}{3/4} + \frac{1}{2}\log\frac{1/2}{1/4} = \frac{1}{2}\log\frac{2}{3} + \frac{1}{2}\log 2 = \frac{1}{2}\log\frac{4}{3} \approx 0.144 \text{ nats.}$$

Pythagorean check. By the Pythagorean theorem for the e-connection: since q lies on the m-geodesic from p to r, the e-geodesic from q is orthogonal to this m-geodesic at q. But p does not in general lie on the e-geodesic from q to some fourth point, so we cannot directly apply the theorem.

Instead, we check the *generalised Pythagorean identity* for the configuration (p, q, r) where q is the m-midpoint:

$$D_{\mathrm{KL}}(p\,\|\,r) = D_{\mathrm{KL}}(p\,\|\,q) + D_{\mathrm{KL}}(q\,\|\,r) + \underbrace{\left(D_{\mathrm{KL}}(p\,\|\,r) - D_{\mathrm{KL}}(p\,\|\,q) - D_{\mathrm{KL}}(q\,\|\,r)\right)}_{\text{remainder}}.$$

Numerical check: $0.082 + 0.144 = 0.226 \neq 0.549 = D_{\mathrm{KL}}(p\,\|\,r)$. The remainder is $0.549 - 0.226 = 0.323 > 0$. The Pythagorean identity holds *only* when the m-geodesic q–r is **G**-orthogonal to the e-geodesic p–q, which is not the case here since p, q, r lie on the same m-geodesic.

Correct application. The Pythagorean theorem $D_{\mathrm{KL}}(p\,\|\,r) = D_{\mathrm{KL}}(p\,\|\,q) + D_{\mathrm{KL}}(q\,\|\,r)$ holds when q is the m-projection of r onto the e-geodesic through p, not when q is on the m-geodesic through p and r. This illustrates the importance of which geodesics are involved. ■

Problem I.3: Natural Gradient for a Gaussian Policy in Reinforcement Learning

Problem. In policy gradient reinforcement learning, the policy is $\pi_\theta(a|s) = \mathcal{N}(\mu_\theta(s), \sigma^2)$ with σ^2 fixed and $\mu_\theta(s) = \theta^\top \phi(s)$ (linear in features $\phi(s)$). Derive the natural gradient of the expected return $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_t r_t]$ with respect to θ, and show that it equals the ordinary gradient preconditioned by the inverse Fisher information matrix of the policy.

Solution. The score function (log-derivative of the policy with respect to θ):

$$\nabla_\theta \log \pi_\theta(a|s) = \frac{a - \mu_\theta(s)}{\sigma^2}\phi(s).$$

Fisher information. For a trajectory $(s_0, a_0, s_1, a_1, \ldots)$:

$$\begin{aligned} I(\theta) &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s)\,(\nabla_\theta \log \pi_\theta(a|s))^\top\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\frac{(a - \mu_\theta(s))^2}{\sigma^4}\phi(s)\phi(s)^\top\right] \\ &= \frac{1}{\sigma^2}\mathbb{E}_{s\sim d^{\pi_\theta}}\left[\phi(s)\phi(s)^\top\right] = \frac{1}{\sigma^2}\Phi^\top D \Phi, \end{aligned}$$

where D is the diagonal matrix of state visitation probabilities under π_θ and Φ is the feature matrix. The term $(a - \mu)^2/\sigma^4$ averages to $1/\sigma^2$ since $\mathrm{Var}_{\pi_\theta}(A|s) = \sigma^2$.

Policy gradient. By the policy gradient theorem:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t)\, Q^{\pi_\theta}(s_t, a_t)] = \frac{1}{\sigma^2}\mathbb{E}[(a_t - \mu_\theta(s_t))\phi(s_t)\, Q(s_t, a_t)]\,.$$

Natural gradient.

$$\tilde{\nabla}_\theta J = I(\theta)^{-1}\nabla_\theta J = \sigma^2(\Phi^\top D\Phi)^{-1} \cdot \frac{1}{\sigma^2}\Phi^\top D\, Q^\pi = (\Phi^\top D\Phi)^{-1}\Phi^\top D\, Q^\pi,$$

where $Q^\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the vector of Q-values. This is precisely the *weighted least-squares solution* of $\Phi w \approx Q^\pi$ with weights D: the natural gradient policy update equals the weighted regression of the Q-function onto the feature space. This is the *compatible function approximation* condition (Sutton et al. 2000), showing that natural policy gradient and compatible value function approximation are equivalent for linear Gaussian policies. ■

Problem I.4: Exponential Family EM for the Beta Distribution

Problem. The Beta distribution $\text{Beta}(\alpha, \beta)$ is an exponential family. Write it in canonical form, identify the sufficient statistics, and derive the EM update for a mixture of K Beta distributions when only the data $x_1, \dots, x_n \in (0, 1)$ is observed (component assignments are latent).

Solution. The Beta density: $p_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$. The log-density:

$$\log p_{\alpha,\beta}(x) = (\alpha - 1)\log x + (\beta - 1)\log(1 - x) + \log\Gamma(\alpha + \beta) - \log\Gamma(\alpha) - \log\Gamma(\beta).$$

This is an exponential family with:

$$\begin{aligned}
\text{Natural parameters: } & \theta = (\theta^1, \theta^2) = (\alpha - 1, \beta - 1) \in (-1, \infty)^2. \\
\text{Sufficient statistics: } & T(x) = (\log x, \log(1 - x)). \\
\text{Log-partition: } & \psi(\theta) = \log\Gamma(\theta^1 + \theta^2 + 2) - \log\Gamma(\theta^1 + 1) - \log\Gamma(\theta^2 + 1). \\
\text{Moment parameters: } & \eta_1 = \mathbb{E}[\log X] = \psi(\alpha) - \psi(\alpha + \beta), \quad \eta_2 = \mathbb{E}[\log(1 - X)] = \psi(\beta) - \psi(\alpha + \beta),
\end{aligned}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function.

Beta mixture EM. For a mixture $p(x) = \sum_{k=1}^K \pi_k\, \text{Beta}(\alpha_k, \beta_k)$ with latent component $Z_i \in \{1, \dots, K\}$:

E-step: Responsibilities $\gamma_{ik} = \pi_k\, p_{\alpha_k,\beta_k}(x_i) / \sum_l \pi_l\, p_{\alpha_l,\beta_l}(x_i)$.

M-step: Update $\pi_k = N_k/n$ where $N_k = \sum_i \gamma_{ik}$, and update (α_k, β_k) by solving the moment-matching equations:

$$\psi(\alpha_k) - \psi(\alpha_k + \beta_k) = \frac{1}{N_k}\sum_i \gamma_{ik}\log x_i, \qquad \psi(\beta_k) - \psi(\alpha_k + \beta_k) = \frac{1}{N_k}\sum_i \gamma_{ik}\log(1 - x_i).$$

These equations must be solved numerically (e.g., by Newton's method using the trigamma function $\psi'(\cdot)$), since the moment map $(\alpha, \beta) \mapsto (\eta_1, \eta_2)$ has no closed-form inverse. The M-step is still an m-projection (moment-matching) in the information-geometric sense; its numerical implementation requires a Newton solve at each iteration. ■

Problem I.5: Fisher-Rao Distance Between Two Gaussians

Problem. Let $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$ be two univariate Gaussians. (a) Write down the Fisher metric and identify the Riemannian geometry of the Gaussian manifold.

(b) Compute the Fisher-Rao geodesic distance $d_{FR}(p, q)$ for the special case $\mu_1 = \mu_2 = 0$. (c) Show that $d_{FR}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\mu + \delta, \sigma^2)) \to |\delta|/\sigma$ as $\delta \to 0$, recovering the signal-to-noise ratio.

Solution.

(*a*) *Fisher metric and geometry.* For $p_{(\mu,\sigma)} = \mathcal{N}(\mu, \sigma^2)$ the log-density is $\log p = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$. The score vector $(\partial_\mu \log p,\, \partial_\sigma \log p) = (\frac{x-\mu}{\sigma^2},\, -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3})$. Taking the covariance under p:

$$\mathbf{G}(\mu, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}, \quad \text{so} \quad ds^2 = \frac{d\mu^2 + 2\, d\sigma^2}{\sigma^2}.$$

The off-diagonal vanishes because $(x - \mu)/\sigma^2$ has mean zero and $((x - \mu)^2/\sigma^2 - 1)/\sigma$ has mean zero, and the cross-term $\mathbb{E}[(x-\mu)((x-\mu)^2-\sigma^2)]/\sigma^5 = 0$ by symmetry of the Gaussian. After the rescaling $\tilde{\sigma} = \sigma\sqrt{2}$, this becomes $ds^2 = (d\mu^2 + d\tilde{\sigma}^2)/(\tilde{\sigma}/\sqrt{2})^2$, the Poincaré upper half-plane metric $ds^2_{\mathbb{H}} = (d\mu^2 + d\tilde{\sigma}^2)/\tilde{\sigma}^2$ with scale factor $\frac{1}{2}$. Thus the Gaussian manifold is isometric to $(\mathbb{H}^2, \frac{1}{2}ds^2_{\mathbb{H}})$, a hyperbolic plane of constant curvature $\kappa = -\frac{1}{2}$.

(*b*) *Distance at equal means* $\mu_1 = \mu_2 = 0$. The path $t \mapsto (0, \sigma(t))$ for $\sigma(t) \in [\sigma_1, \sigma_2]$ $(\sigma_1 < \sigma_2)$ has length

$$d_{FR}(\mathcal{N}(0, \sigma_1^2), \mathcal{N}(0, \sigma_2^2)) = \int_{\sigma_1}^{\sigma_2} \sqrt{g_{\sigma\sigma}}\, d\sigma = \int_{\sigma_1}^{\sigma_2} \frac{\sqrt{2}}{\sigma}\, d\sigma = \sqrt{2}\log\frac{\sigma_2}{\sigma_1}.$$

This is the geodesic distance along the imaginary axis of the Poincaré upper half-plane (the unique Levi-Civita geodesic for the hyperbolic metric that runs vertically), confirming that scale differences are measured logarithmically in the Fisher-Rao geometry.

(*c*) *Infinitesimal limit.* The Fisher-Rao length element in the μ-direction at fixed σ is $\sqrt{g_{\mu\mu}}\,|d\mu| = (1/\sigma)|d\mu|$. Therefore $d_{FR}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\mu + \delta, \sigma^2)) \approx |\delta|/\sigma$ for small $|\delta|$. The ratio $|\delta|/\sigma$ is the classical signal-to-noise ratio: two Gaussian distributions with the same variance σ^2 are hard to distinguish (large σ, small d_{FR}) when the mean difference $|\delta|$ is much smaller than σ, and easy to distinguish (large d_{FR}) when $|\delta| \gg \sigma$. The Fisher-Rao metric makes this operational: it quantifies *statistical distinguishability* rather than parameter distance. ■

Problem I.6: Natural Gradient for a Categorical Distribution and the Softmax Connection

Problem. Let p_θ be the K-categorical distribution with natural parameters $\theta = (\theta_1, \dots, \theta_K)$ subject to $\sum_k \theta_k = 1$ (i.e. θ is a probability vector, so we use the mean parametrisation $\eta = \theta$ directly). (a) Compute the Fisher matrix $\mathbf{G}(\theta)$ and its inverse. (b) Let $L(\theta) = -\sum_k y_k \log \theta_k$ be the cross-entropy loss for a one-hot target $y = (y_1, \dots, y_K)$ (a single observed class c, so $y_k = \mathbb{1}[k = c]$). Compute the natural gradient $\tilde{\nabla} L = \mathbf{G}^{-1}\nabla L$ and interpret it. (c) Compare natural gradient descent on L with ordinary gradient descent, and show that one step of natural gradient descent (with appropriate step size) equals an exact Newton step in this setting.

Solution.

(*a*) *Fisher matrix.* For $p_\theta(k) = \theta_k$ the score is $\partial_j \log p_\theta(k) = \mathbb{1}[k = j]/\theta_j$. The Fisher matrix is

$$\mathbf{G}_{ij}(\theta) = \mathbb{E}_\theta\left[\frac{\mathbb{1}[X = i]}{\theta_i} \cdot \frac{\mathbb{1}[X = j]}{\theta_j}\right] = \frac{\mathbb{1}[i = j]}{\theta_i}.$$

Thus $\mathbf{G}(\theta) = \mathrm{diag}(1/\theta_1, \dots, 1/\theta_K)$ and $\mathbf{G}(\theta)^{-1} = \mathrm{diag}(\theta_1, \dots, \theta_K)$.

However, the simplex constraint $\sum_k \theta_k = 1$ makes θ a point in an $(K - 1)$-dimensional manifold, not the full $\mathbb{R}^K$. On the constrained manifold the projected Fisher matrix is $\mathbf{G}_{ij} =$

$\theta_i\delta_{ij} - \theta_i\theta_j$ (the multinomial covariance matrix), and its Moore-Penrose pseudoinverse is $\mathbf{G}^+_{ij} = \delta_{ij}/\theta_i - 1$ (the correction subtracts the mean). For the unconstrained (log-sum-exp softmax) parametrisation the diagonal formula holds exactly.

(*b*) *Cross-entropy natural gradient.* With $L(\theta) = -\log\theta_c$ (cross-entropy for class c):

$$\nabla_j L = -\frac{\mathbb{1}[j=c]}{\theta_c} = -\frac{y_j}{\theta_j}.$$

The natural gradient in the unconstrained diagonal-Fisher setting:

$$(\tilde{\nabla}L)_j = (\mathbf{G}^{-1}\nabla L)_j = \theta_j \cdot \left(-\frac{y_j}{\theta_j}\right) = -y_j.$$

The natural gradient is *exactly the negative one-hot vector*: $\tilde{\nabla}L = -y$. This is remarkably clean: regardless of the current prediction θ, the natural gradient always points in the direction $-y$ (increase the probability of the true class by 1, decrease others by 0 in the natural direction). The Fisher matrix has *exactly cancelled* the $1/\theta_j$ scaling of the ordinary gradient.

(*c*) *Newton step comparison.* The Hessian of L in mean parametrisation is $\nabla^2 L_{jk} = \delta_{jk}\cdot\mathbb{1}[j=c]/\theta_c^2 = \delta_{jk}y_j/\theta_j^2$. The Newton step is $-(\nabla^2 L)^{-1}\nabla L$: for component $j = c$, $(-1/\theta_c^2)^{-1}\cdot(-1/\theta_c) = \theta_c$; for $j \neq c$, 0. So the Newton update on θ_c is $+\theta_c$, giving $\theta_c^{\text{new}} = \theta_c + \theta_c = 2\theta_c$ (then renormalize), which moves θ_c toward 1 but is not the same as the natural gradient.

In the *natural* (exponential family) parametrisation $\phi_k = \log\theta_k - \log\theta_K$ (the softmax parametrisation), the Fisher matrix becomes the multinomial covariance $\tilde{\mathbf{G}}_{ij}(\phi) = \theta_i\delta_{ij} - \theta_i\theta_j$, and a single natural gradient step from the MLE first-order condition $\nabla_\phi L = \theta - y = 0$ solves the problem exactly in one step — the natural gradient is exactly the Newton step in the canonical exponential family coordinate system. This is the information-geometric justification for why cross-entropy training of softmax classifiers is well-behaved: in natural coordinates, the loss is a quadratic function of the parameter (modulo the log-partition function), and the natural gradient descent is a Newton method. ■